Design effects: model-based versus design-based approach

Ganninger, Matthias

Veröffentlichungsversion / Published Version
Dissertation / phd thesis

gesis
Leibniz-Institut
für Sozialwissenschaften

Schriftenreihe Band 3

Design Effects:

Model-based versus Design-based Approach

Matthias Ganninger

Design Effects: Model-based versus Design-based Approach

GESIS-Schriftenreihe
herausgegeben von GESIS – Leibniz-Institut für Sozialwissenschaften

Band 3

Matthias Ganninger

Design Effects:
Model-based versus Design-based Approach

Die vorliegende Arbeit wurde vom Fachbereich IV, Wirtschafts- und Sozialwissenschaften, Mathematik, Informatik und Wirtschaftsinformatik der Universität Trier im Jahr 2009 als Dissertation zur Erlangung des akademischen Grades eines Doktors der Wirtschafts- und Sozialwissenschaften (Dr.rer.pol.) angenommen.

Erstgutachter: Prof. Dr. Ralf Münnich
Zweitgutachter: PD Dr. Siegfried Gabler
Tag der Disputation: 3. Dezember 2009
Vorsitzender: Prof. Dr. Christian Bauer

Matthias Ganninger

Design Effects:

Model-based versus Design-based Approach

GESIS – Leibniz-Institut für Sozialwissenschaften 2010

Bibliographische Information Der Deutschen Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

ISBN 978-3-86819-010-6
ISSN 1869-2869

Herausgeber,
Druck u. Vertrieb: GESIS – Leibniz-Institut für Sozialwissenschaften
Lennéstraße 30, 53113 Bonn, Tel.: 0228 / 22 81 -0
info@gesis.org
Printed in Germany

Contents

List of Symbols

PSU Level – Population Quantities

$$M = \text{number of PSUs in the population,}$$

$$N_i = \text{number of SSUs in the } i\text{th PSU in the population,}$$

$$N = \sum_{i=1}^{M} N_i = \text{total number of SSUs in the population}$$

$$\bar{B} = \frac{N}{M} = \text{mean PSU size in the population,}$$

$\bar{B} = N_i, \forall i$ if all PSUs are of the same size

$$t_{iU} = \sum_{j=1}^{N_i} Y_{ij} = \text{population total of the } i\text{th PSU}$$

$$t_U = \sum_{i=1}^{M} t_{iU} = \sum_{i=1}^{M}\sum_{j=1}^{N_i} Y_{ij} = \text{population total}$$

$$\bar{Y}_U = \frac{t_U}{N} = \text{population mean}$$

$$S^2_{t_U} = \frac{1}{M-1}\sum_{i=1}^{M}\left(t_{iU} - \frac{t_U}{M}\right)^2 = \text{population variance of the PSU totals}$$

SSU Level – Population Quantities

$$\bar{Y}_{iU} = \frac{1}{N_i}\sum_{j=1}^{N_i} Y_{ij} = \text{population mean of the } i\text{th PSU,}$$

$j = 1, \ldots, N_i$

$$\bar{Y}_U = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N_i} Y_{ij} = \text{population mean}$$

$$S^2_{iU} = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}\left(Y_{ij} - \bar{Y}_{iU}\right)^2 = \text{population variance of the } i\text{th PSU}$$

$$S^2_U = \frac{1}{N-1}\sum_{i=1}^{M}\sum_{j=1}^{N_i}\left(Y_{ij} - \bar{Y}_U\right)^2 = \text{population variance}$$

Sample Quantities

m = number of PSUs in the sample s

n_i = number of SSUs in the sample of the ith PSU

$n = \sum_{i \in s^{(1)}} n_i$ = total number of SSUs in the sample

$\bar{b} = \frac{n}{m}$ = mean PSU size in the sample

π_i = inclusion probability of the ith PSU

$\pi_{j|i}$ = inclusion probability of the jth SSU given the ith PSU was selected

$\pi_i \pi_{j|i}$ = overall probability that the (i, j)th element is selected

$w_i = \frac{1}{\pi_i}$ = design weight of the ith PSU

$w_{ij} = \frac{1}{\pi_i \pi_{j|i}}$ = overall design weight of the (i, j)th element

$\bar{y}_i = \frac{1}{n_i} \sum_{j \in s_i} y_{ij}$ = sample mean of the ith PSU

$\bar{y}_s = \sum_{i \in s} \bar{y}_i$ = sample mean

$\hat{t}_i = \frac{M_i}{n_i} \sum_{j \in s_i} y_{ij}$ = estimated total of Y in the ith PSU (srs on stage two)

$\hat{t}_{unb} = \sum_{i \in s^{(1)}} \frac{\hat{t}_i}{\pi_i}$ = unbiased estimator of the population total

$s_t^2 = \frac{1}{m-1} \sum_{i \in s} \left(\hat{t}_i - \frac{\hat{t}_{unb}}{m} \right)^2$ = estimated variance of PSU totals (srs on both stages)

$s_i^2 = \frac{1}{n_i - 1} \sum_{j \in s_i} \left(y_{ij} - \bar{y}_i \right)^2$ = sample variance within the ith PSU (srs on stage two)

List of Tables and Figures

Preface

I would like to thank my supervisors, Ralf Münnich and Siegfried Gabler for their enduring support and encouragement. Without their enlightening and inspiring guidance, this thesis would not have been possible. The support of Sabine Häder, a dear colleague of mine at GESIS, often helped me gain new insights. Discussions with Partha Lahiri during his visits to GESIS always were inspiring and often laid the ground for new pathways to further reaching research questions.

My sincere thanks also goes to all other colleagues at GESIS who supported me during the time when this thesis was written. Among those, Annelies Blom and Dorothée Behr receive a special thanks.

Last but not least, I am deeply thankful for the great amount of patience and good will which my parents showed me during the last years. Together with my grand mother they served as a constant source of motivation all along this winding road.

1 Introduction

The need for adequate consideration of the effects of a sample design on the precision of estimators is becoming recognized by an increasing number of sample survey projects like TIMMS (Gonzalez and Foy, 2004), the URGE study (Campbell et al., 2000), the Add Health study (Chantala and Tabor, 1999), population-based diarrhea prevalence surveys (Katz et al., 1993), general health surveys like Health-2000 (Lehtonen et al., 2002), epidemiological studies like PAQUID (Lemeshow et al., 1998), nutrition examination surveys (Lago et al., 1987b) or crime victimization studies (Schnell and Kreuter, 2005) to name only a few. These studies recognize the need to take into account the possibly negative effects of a complex sample design on estimators of interest. They make use of a concept called *design effect* which can serve as a measure of variance inflation in the estimator due to a departure from simple random sampling.

The European Social Survey (ESS) was the first general social survey to make explicit use of design effects already at the planning stage (ESS, 2005a). In this biennial, EU-wide general social survey each participating country is responsible for its sample to meet some pre-defined quality criteria. One of these criteria concerns the precision of estimators: The samples of all participating countries shall yield estimators of comparable precision (ESS, 2005b, 1). Design effects play a crucial role in the planning of samples which will yield estimators with these properties.

The foundations for sampling in Europe-wide surveys like the ESS are, however, quite diverse. In some countries, like Sweden, Norway or Finland, scientific researchers are allowed to draw a sample directly from population registers. In other countries, like Portugal, Spain or Poland, access to population registers is either limited or not possible at all. With this diversity of sampling frames comes a diversity of sample designs. Whereas in countries of the first group, a simple random sample (srs), a stratified random sample (str) or a systematic sample (sys) of contact persons can be drawn directly, this is not possible in the second group of countries due to the structure of the sampling frame. Often, these countries have to resort to more complex sample designs like for example cluster or multi-stage sample designs. In a multi-stage sample design one draws, for example, municipalities at the first stage. Then, at the second stage, persons are drawn from a complete list of inhabitants within each municipality. It is an empirical fact, however, that persons who are socialized within the same social context (e.g. living in the same neighborhood or municipality), are more similar to each other than to persons who are socialized in another social context on many queried items of a general social survey like the ESS. This *homogeneity* can have a negative effect on the precision of estimators.

The accuracy of an estimator calculated with data which have arisen from a simple random sample differs from the quality of the same estimator calculated on the basis of a cluster sample design described above, given the two samples are of the same size. Nevertheless, all samples in the ESS have to comply with the aforementioned quality standards in terms of the precision of estimators. The question, then, is how to

design different samples such that these criteria are met under the practical restriction of divergent sampling frames.

One way to ensure that the precision of an estimator is independent of the sample design is to plan samples with an equal *effective sample size*. The effective sample size is a concept which incorporates the design effect. A thorough formal definition of the design effect follows in Chapter 3. For the time being, the design effect shall be defined as a measure for the inflation of variance of an estimator under a given (e.g. cluster or multi-stage) sample design compared to the variance of the same estimator under simple random sampling. There is, however, no unique design effect for a given sample. Design effects will vary in magnitude depending on the characteristics of the item under study. The following example illustrates the connection between a) a study variable, b) the definition of clusters and c) the effects of different sample designs (Kish, 1989).

Let us thus assume a clustered population in which values of the study variable are distributed according to Table 1. The column and row means ($\bar{y}_{\bullet c}$ and $\bar{y}_{r\bullet}$, respectively) are given along with the variable values.

Table 1: *A clustered population*

						$\bar{y}_{r\bullet}$
	1	6	11	16	21	11
	2	7	12	17	22	12
	3	8	13	18	23	13
	4	9	14	19	24	14
	5	10	15	20	25	15
$\bar{y}_{\bullet c}$	3	8	13	18	23	

From this population $n = 10$ elements are to be drawn by a) srs and b) cluster sampling where clusters are either defined by columns (clu-col) or rows (clu-row) of the matrix. Under cluster sampling all elements of two randomly selected columns or rows are selected. Under srswr n elements are chosen randomly. In either case the sample mean $\bar{y} = n^{-1} \sum_{i=1}^{n} y_i$ is to be calculated. The means and the variances under srs, $Var_{(\text{srswr})}(\bar{y})$, under column-wise cluster sampling, $Var_{(\text{clu}-\text{col})}(\bar{y})$ and under row-wise cluster sampling, $Var_{(\text{clu}-\text{row})}(\bar{y})$, are given in the following table.

	srswr	clu-col	clu-row
Mean	13.00	13.00	13.00
Variance	5.20	18.75	0.75

It can be seen that the population mean of $\bar{Y} = 13$ is estimated without bias under all sample designs but the variances of the estimates vary dramatically. The variance of the estimates of $\bar{y}$ under srswr (5.2) shall serve as a reference. Under column-wise cluster sampling the variance of the sample mean is 18.75 which is $\frac{Var_{(\text{clu}-\text{col})}(\bar{y})}{Var_{(\text{srswr})}(\bar{y})} = \frac{18.75}{5.2} = 3.61$ times $Var_{(\text{srswr})}(\bar{y})$. If the columns selected are not exactly symmetrical to the third column (i.e. first and fifth and second and fourth), the difference between the sample mean and the population parameter will be very large. If columns are

sampled, the variance of sample mean is very low $\left(\frac{Var_{(\text{clu-row})}(\bar{y})}{Var_{(\text{srswr})}(\bar{y})} = \frac{0.75}{5.2} = 0.14\right)$. This is due to the very low heterogeneity of row-wise means. Even if, in one of the worst cases, for example the two upper rows are selected, the sample mean is 11.5 which is closer to the population mean than in one of the corresponding worst cases of column-wise selection (i.e. if for example the first or last two columns are selected) where the sample mean is 5 and 20.5, respectively.

This example illustrates that cluster sampling can yield better and worse results (in terms of precision) than srswr. The magnitude of loss or gain in precision depends on the interrelation of the distribution of the study variable and the structure and definition of clusters in a sample design. In most real-world sample surveys, however, these two parameters are interrelated in a way such that precision is lost.

If design weights are constant and no additional inflation of variance is generated through weighting, the only factor by which the variance is underestimated is due to clustering. For sake of simplicity, the additional inflation of variance due to unequal inclusion probabilities will be introduced later. For the time being, everything that follows is discussed under the assumption of equal inclusion probabilities.

For a given sample and a certain item, design effects can be estimated under two different approaches: the design-based approach to design effects is flexible in terms of the estimators for which *deff* can be specified. The model-based approach has the advantage of enabling predictions of expected design effects. The components of the estimators of *deff* under either approach are, however, subject to quality issues themselves. Thus, the question arises which estimation approach yields the best results in terms of bias and precision. In the following this question shall be answered for a set of practically relevant estimators and sample designs.

To arrive at an answer, Chapter 2 will first give some basic definitions and clarifications of notation and terminology. Then, Chapter 3 introduces the concept of design effects more stringently and gives an overview of estimation techniques. In Chapter 4, measures of homogeneity are introduced and their merits and shortcomings are discussed. Since some of these measures of homogeneity are based on the variance components of a random effects model, Chapters 7.3 and 7.3 give a basic introduction into linear models, generalized linear models and random effect models. In Chapter 5, a Monte Carlo simulation study investigates the behavior of estimators of the design effect and its components. Similar investigations are conducted on the basis of selected countries of the ESS in Chapter 6. The findings of both the Monte Carlo study and those based on the ESS are synthesized and discussed in Chapter 7.

2 Technical Foundations

In this chapter the technical foundations underlying the most frequently used concepts of this thesis are introduced. It proceeds with a definition of the investigated sample designs (Section 2.1). Section 2.2 defines estimators and their variance estimators under study. Then, Section 2.3 gives an introduction into some of the elementary aspects of design effects. Finally, Section 2.4 illustrates the effects of a complex sample design on the variance of an estimator.

2.1 Sample Designs

Let U denote a universe of size N such that $U_1, \ldots, U_i, \ldots, U_N$ are the elements of U. A *study variable* is denoted by Y and has unknown population values $Y_1, \ldots, Y_i, \ldots, Y_N$. A *sample*, s, is a subset of U. The set of all possible samples of size n is denoted by S. A sample selection scheme is a mechanism which assigns a specific sample, s, a non-zero selection probability, $p(s)$. Generally, the function $p(\cdot)$ is called *sample design* or *sampling scheme*. Sample designs which do explicitly define a function $p(\cdot)$ are called *probability sample designs*. Sample designs which do not explicitly define a function $p(\cdot)$ are called *non-probability sample designs* and are not considered here.

For a given sample design, each element of the population has an *inclusion probability*, π_i, which indicates the a priori probability of the element to be selected into the sample. The inverse of π_i is called *design weight* and is defined as $w_i = \pi_i^{-1}$. If sampling is *with replacement* (wr), each population element can occur more than once in the sample since it becomes a member of U again once it has been selected. If an element of the population does not have a chance of occurring in the sample more than once, the sample design is *without replacement* (wor). The study variable, Y, is surveyed at each element of the sample. Values of Y in the sample are denoted by $y_1, \ldots, y_i, \ldots, y_n$. The elements of s are called *ultimate sampling units* if y can be surveyed directly on them. A *sample survey* is the combination of the realization of a sample design and measurement of the values of at least one study variable for elements of the sample.

A sample design is called *simple random sample* (srs) if all $s \in S$ have the same selection probability. A *complex sample design* is any probability sample design which is not srs. Although complex sample designs suffer from significant problems, they represent the most commonly applied sampling scheme in the social sciences and related fields of research (Lohr, 1999, pp. 221). Their popularity is based on a number of practical advantages associated with each design's specific characteristics. Some of these advantages and shortcomings are discussed in more detail in Subsections 2.1.1 and 2.1.2.

Following the delineation in Schnell et al. (1999), sample designs can further be classified as depicted in Figure 1[1]. Of the sample designs presented in figure 1, *multistage sampling*, *cluster sampling*, *disproportional stratified sampling*, and combina-

1 In addition to simple random sample designs mentioned in the box at the bottom of the illustration, systematic and πps designs also play an important role in survey sampling.

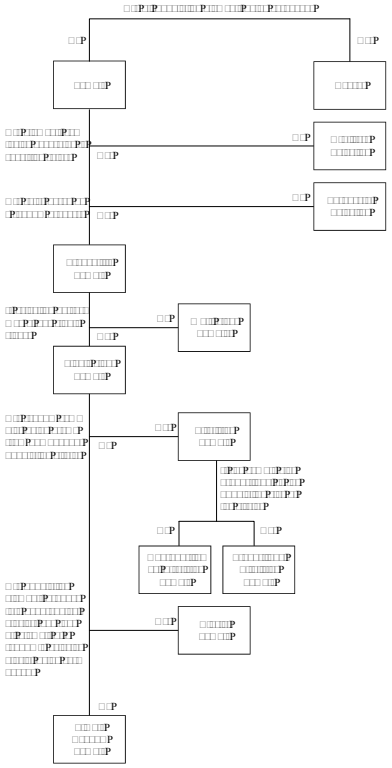


Figure 1: *Sampling Schemes; in Schnell et al. (1999, 252)*

tions of these are at the focus of analysis. Deviating from Figure 1, the term *cluster sampling* will be used more widely and will also refer to multi-stage sample designs since any multi-stage sample design possesses elements of clustering at one or more stages.

As mentioned above, these sample designs shall be referred to as belonging to the class of complex sample designs. Sample surveys which apply any of these designs or combinations of them are called *complex sample surveys*. There exist more sampling schemes which are also referred to as complex sample designs (e.g. systematic sampling, Sampford sampling, etc.). For the purpose of this thesis, however, the term will be defined more narrowly. Put more stringently, complex sample designs differ from srs in at least one of the following aspects: a) inclusion probabilities, b) stratification, c) clustering.

In complex sample designs inclusion probabilities may vary – either because the researcher intentionally plans to over-sample a specific sub-group or due to other reasons associated with the sample design. Sample designs which cause inclusion probabilities to vary are called *unequal probability* sample designs. In the following

subsections, two important complex sample designs are introduced in more detail: *stratification* on the one hand and multi-stage sampling on the other hand as these are the most prominent and widely used design modifications. Stratification is usually applied when the sampling frame offers information on a characteristic, Z, useful for stratification. Sensible stratification can reduce sampling error and hence increase the precision of estimators. A multi-stage sample, on the other hand, is either drawn because no frame of ultimate sampling units exists or because fieldwork personnel management aims at a geographical allocation of the interviewers that minimizes travelling between and within geographical clusters. Clustering can, however, introduce a severe loss in precision of estimators.

2.1.1 Cluster Sampling and Multi-Stage Sampling

A *cluster sample design* (clu) is any sample design in which ultimate sample units are not selected directly from a frame but from a sample of superordinate non-overlapping *clusters*. A *cluster*, or primary sampling unit (*PSU*), denotes a subset of population units which belong to this subset due to some well defined specific (known or unknown) attributes (e.g. a respondent's address in the case of geographical clustering or the person a respondent is interviewed by if the survey is conducted face-to-face or by telephone).

Each ultimate sample element belongs to exactly one PSU and each cluster is comprised of one or more ultimate sampling elements. A clustered population consists of M PSUs, each of the same size $N_i = C, i = 1, \ldots, M$. We shall assume there exists a complete frame of PSUs from which a sample of m PSUs is drawn. The set of possible samples of m of the M clusters is denoted by S and a specific sample of m PSUs is denoted by s. The cluster sample design is defined by $p(s)$. The inclusion probabilities of each of the M clusters is denoted by π_i with $i = 1, \ldots, M$. The value of π_i depends on the characteristics $p(s)$.

After s has been obtained, y is being surveyed for each of the $n = m \times C$ ultimate sample elements (ignoring contact and non-response issues for the time being). This sampling scheme is referred to as *single-stage sampling* or simply *cluster sampling*. There exist, however, a wide range of variations to this most simple clustered sample design.

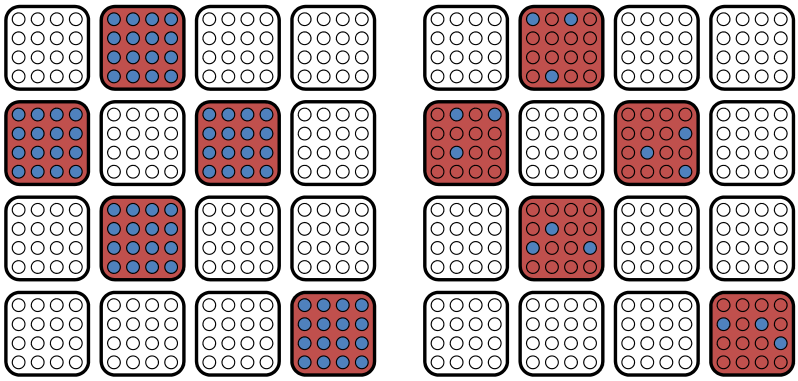
Multi-stage sampling is any sample design in which ultimate sample elements are selected through subsequent sampling on two or more superordinate stages.

In *two-stage sampling* (clu2 for short), for example, m of M clusters are selected at the first stage. The set of possible samples of m primary sampling units is denoted by $S^{(1)}$. A specific sample of m primary sampling units is denoted by $s^{(1)}$, inclusion probabilities for each of the M PSUs are denoted by π_i, $i = 1, \ldots, M$.

At the second stage, n_i *secondary sampling units* (*SSU*) of the ith PSU of size $N_i = C$ are selected for $i \in s^{(1)}$. Thus, $n = \sum_{i \in s^{(1)}} n_i$. Elements of the ith cluster are denoted by $1, \ldots, j, \ldots, n_i$. The set of possible samples of n_i from N_i SSUs in the ith PSU is denoted by $S_i^{(2)}$ and a specific sample by $s_i^{(2)}$.

The inclusion probability of the jth element given the ith PSU was selected is $\pi_{j|i}$ and the design weight for the (i,j)th element is given by $w_{ij} = \pi_{ij}^{-1}$. A consistent notation is used in three- or more generally in multi-stage sampling.

Two-stage sampling is the sample design which underlies most of the empirical analyses in chapter 5 and an important design also in the ESS (as discussed in chapter 6). Figure 2 gives examples of single-stage cluster sampling (cluster sampling) with $M = 16$, $N_i = 16$ and $m = 5$ and two-stage cluster sampling with $m = 5$ and $n_i = 3$.



(a) One-stage cluster sampling with $m = 5$

(b) Two-stage cluster sampling with $m = 5$ and $n_i = 3$

Figure 2: *One-stage and two-stage cluster sampling scheme*

It is a widely spread believe that one of the most striking advantages of cluster sampling for social surveys is that it guarantees reduced travel costs. Interviewers can be sent to the field within closely defined geographical boundaries. Often, a primary sampling unit is defined as a municipality or a city district making travelling from address to address relatively inexpensive. We will later see that if the estimators based on data of a geographically clustered sample design are estimated naively, this assumption may hold but that it can be neglected if the effects of the sample design are incorporated in the estimation process.

A further explanation for the wide spread use of cluster sample designs are the admission restrictions of alternative sampling frames of ultimate sample elements (e.g. through population registers). In fact, many European countries either lack of such a list or do not allow researchers to draw a sample from it. This is also reflected in the sample designs used by ESS countries from which only about half are not multi-stage designs (ESS, 2005b; Häder et al., 2007). In the ESS, the guideline for selection of a sample design follows the recommendation of Kish (1994, 173): "Sample designs may be chosen flexibly and there is no need for similarity of sample designs. Flexibility of choice is particularly advisable for multinational comparisons, because the foundations for samplig differ greatly between countries. All this flexibility assumes

probability selection methods: known probabilities of selection for all population elements" .

Thorough definitions of the underlying concepts can be found in the list of symbols on page 9.

2.1.2 Stratified Sampling

In stratified sampling (str), the population of interest is divided into H non-overlapping *sub-populations* or *strata* of size $N_h, h = 1, \ldots, H$. The set of possible samples of n_h from N_h is denoted by S_h and a specific sample by s_h. The sample design $p(s_h)$. A sample of size n_h is drawn from each of the H strata by $p(s_h)$ such that $n = \sum_{h=1}^{H} n_h$.

Basically, there are two types of mechanism to obtain n_h. First, n_h can be allocated proportionally to the population figure N_h such that $n_h = n \times \frac{N_h}{N}$. The resulting sampling scheme is called *proportional stratified sampling* (strp). On the other hand, any stratified sample that is not according to that allocation method is called *disproportional stratified sample* (strd).

Stratified samples can have a lower variance than srs. The magnitude of reduction or increase of variance, however, depends on the degree of homogeneity of elements within the strata and heterogeneity between strata. Thus, a well-informed choice of stratification characteristics is essential to achieve the promising gains in efficiency that stratification generally offers. For a more detailed overview of stratification techniques the reader must be referred to Särndal et al. (1992, chapter 3.7), Cochran (1977), Lehtonen and Pahkinen (2004, pp. 61) or Münnich (2003b).

2.2 Estimators

A *population parameter*, θ, is a function of the values of the study variable. The population total of Y, for example, is given by $t_U = \sum_{i=1}^{M} \sum_{j=1}^{N_i} Y_{ij}$ and the population mean is given by $\bar{Y}_U = \frac{t_U}{N}$. An *estimator*, $\hat{\theta}$, is a function of the observed values of the study variable. An *estimate* is the numeric value produced by an estimator. An *unbiased estimator* has the property that its expected value equals the population parameter, $E\left(\hat{\theta}\right) = \theta$.

In unequal probability sampling the *Horvitz-Thompson estimator* (*HT estimator*) is an unbiased estimator of the population total and is defined as

$$\hat{t}_{\mathrm{HT}} = \sum_{i \in s} \frac{y_i}{\pi_i} = \sum_{i \in s} w_i y_i \qquad , \tag{2.1}$$

where $w_i = 1/\pi_i$ is the design weight of the ith element (see Lohr, 1999, 207). The HT estimator of the population mean is given by

$$\hat{\bar{y}}_{\mathrm{HT}} = \frac{\hat{t}_{\mathrm{HT}}}{N} \qquad . \tag{2.2}$$

The combination of an estimator and a sample design $p(\cdot)$ is called *strategy*. The following is, with minor modifications, closely leant on Lohr (1999, pp. 196).

The *variance of an estimator* is generally denoted by $Var(\hat{\theta})$. As a rule, those estimators with small variances are preferred over estimators with large variances. The variance of one estimator under one sample design can differ from the variance of the same estimator under another sample design. This why we shall write, more generally, $Var_{p(\cdot)}\left(\hat{\theta}\right)$ to take into account the dependence on the sample design. A *variance estimator* is an estimator for the variance of an estimator and is generally denoted by $\widehat{Var}_{p(\cdot)}\left(\hat{\theta}\right)$. The variance of the HT estimator for the population total is given by

$$Var_{p(s)}\left(\hat{t}_{\mathrm{HT}}\right) = \sum_{i=1}^{N} \frac{1-\pi_i}{\pi_i} t_i^2 + \sum_{i=1}^{N} \sum_{q \neq i}^{N} \frac{\pi_{iq} - \pi_i \pi_q}{\pi_i \pi_q} t_i t_q + \sum_{i=1}^{N} \frac{Var(\hat{t})}{\pi_i} \quad , \tag{2.3}$$

if $|s| = n$ for $p(s) > 0$ and π_{iq} is the probability that both elements i and q are in the sample. The above formula can also be expressed in the so called Sen-Yates-Grundy form as

$$Var_{p(s)}^{(\mathrm{SYG})}\left(\hat{t}_{\mathrm{HT}}\right) = \sum_{i=1}^{N} \sum_{q>i}^{N} \left(\pi_i \pi_q - \pi_{iq}\right) \left(\frac{t_i}{\pi_i} - \frac{t_q}{\pi_q}\right)^2 + \sum_{i=1}^{N} \frac{Var(\hat{t})}{\pi_i} \quad . \tag{2.4}$$

The estimated variance of the HT estimator for the population total is then given by

$$\widehat{Var}_{p(s)}\left(\hat{t}_{\mathrm{HT}}\right) = \sum_{i \in s} \left(1-\pi_i\right) \frac{\hat{t}_i^2}{\pi_i^2} + \sum_{i \in s} \sum_{\substack{q \in s \\ q \neq i}} \frac{\pi_{iq} - \pi_i \pi_q}{\pi_{iq}} \frac{\hat{t}_i}{\pi_i} \frac{\hat{t}_q}{\pi_q} + \sum_{i \in s} \frac{\widehat{Var}_{p(s)}\left(\hat{t}_i\right)}{\pi_i} \tag{2.5}$$

and, correspondingly, in Sen-Yates-Grundy form by

$$\widehat{Var}_{p(s)}^{(\mathrm{SYG})}\left(\hat{t}_{\mathrm{HT}}\right) = \sum_{i \in s} \sum_{\substack{q \in s \\ q > i}} \frac{\pi_i \pi_q - \pi_{iq}}{\pi_{iq}} \left(\frac{\hat{t}_i}{\pi_i} - \frac{\hat{t}_q}{\pi_q}\right)^2 + \sum_{i \in s} \frac{\widehat{Var}_{p(s)}\left(\hat{t}_i\right)}{\pi_i} \quad . \tag{2.6}$$

Formulas for the variance estimator of HT estimator for the population mean can be found in Kish (1965, pp. 255). Estimation of the variance in this closed form requires computation of second order inclusion probabilities, π_{iq}, which can become cumbersome or impossible at all as Münnich (2008, 322) notes. Thus, practical approximations have to be found that avoid use of second order inclusion probabilities. One such approximation for one-stage fixed n sample designs was suggested by Deville (1999) and is given by

$$\widehat{Var}_{p(s)}^{(\mathrm{Deville})}\left(\hat{t}_{\mathrm{HT}}\right) = \frac{1}{1 - \sum_{i \in s} a_i^2} \sum_{i \in s} \left(1-\pi_i\right) \left(\frac{y_i}{\pi_i} - \sum_{j \in s} a_j \frac{y_j}{\pi_j}\right)^2 \quad , \tag{2.7}$$

where $a_i = \dfrac{1-\pi_i}{\sum_{j\in s}\left(1-\pi_j\right)}$.

In addition, Section 5.4.2 will present results on estimating the design effect for the median. Let the study variable in the population of size N follow the distribution function $F(\cdot)$ and let it be ordered such that $y_1 \leq \cdots \leq y_N$. The median, $\tilde{y}$, for y-values with a smooth distribution function is defined as $F(\tilde{y}) = \frac{1}{2}$ (Särndal et al., 1992, 197). An estimator of the median based on a sample of size n is given by

$$\hat{\tilde{y}} = \begin{cases} y_{z+1} & \text{if } n = 2z+1 \text{ (i.e. } n \text{ is odd)} \\ \frac{y_z + y_{z+1}}{2} & \text{if } n = 2z \text{ (i.e. } n \text{ is even)} \end{cases} . \tag{2.8}$$

2.3 Elements of Design Effects

A complex (i.e. cluster or multi-stage) sample design often leads to an inflation of the variance of an estimator only to the degree to which the population (and hence, the sampled elements) show some homogeneity within the same clusters and less homogeneity between clusters. Put formally, values of the study variable follow the model M1 (Gabler et al., 2006, 4)

$$\begin{aligned} Var\left(y_{ij}\right) &= \sigma^2 \\ Cov\left(y_{ij}, y_{i'j'}\right) &= \begin{cases} \sigma^2\rho & \text{for } i = i', j \neq j' \\ 0 & \text{otherwise.} \end{cases} \end{aligned} \tag{2.9}$$

The above model implies that the variance of the study variable is inflated by the factor of ρ for all elements of the same cluster. A very simple example for an item which follows that model is made up by the following question: "*How many kilometres away from the Dome of Cologne do you live?*" If this question is asked in a multi-stage sample conducted in Germany with cities or city blocks as primary sampling units and a fixed number of respondents sampled within each PSU, the answers of the respondents of each PSU would be almost exactly the same. Thus, a large share of the total variation on that item comes from between PSUs and hardly any can be attributed to variation within clusters. For reasons of simplicity, assume that in fact all respondents within each PSU give *exactly* the same answer. Hence, the gain in information after having surveyed the answer of the first respondent in a PSU is zero because the second, the third and all other persons of the cluster will give exactly the same answer.

Assume that $m = 200$ PSUs have been sampled and that $n_i = 10, i = 1, \ldots, m$, persons in each PSU have been asked the above question. However, there are only 200 unique answers to the item which contribute to the variance of that variable due to the lack of independence of the answers. Thus, when falsefully treating the sample data as fulfilling the assumption of independence and using, for example, the usual variance estimator of the sample mean, the true variance (i.e. the variance

of an appropriate estimator not assuming independence of the responses of the study variable) of the estimator will be underestimated by a factor which is called the *design effect*. The design effect depends on the average size of clusters, on the extent to which respondents of a certain cluster resemble each other as well as on the dissimilarity of respondents belonging to different clusters.

Although many empirical studies show that design effects are anything but negligible (Bieler and Williams, 1990; Gabler and Häder, 2000; Gonzalez and Foy, 2004; Kish, 1995; Lemeshow et al., 1998; Lehtonen et al., 2002; Rowe et al., 2002; Selhub et al., 1999; Verma et al., 1980; Yeo et al., 1999), the concept is commonly ignored in substantive analyses. Including design effects in the data analysis usually yields more conservative results for example in hypothesis tests (i.e. a significant difference may no longer be significant if design effects are taken into account). This may be one reason why design effects are widely neglected.

Besides the homogeneity introduced by spatial clustering, homogeneity may also be due to interviewers. This *interviewer effect* follows the same logic as the design effect and can be interpreted similarly: Respondents surveyed by the same interviewer may form a kind of artificial cluster – being interviewed by the same interviewer may make their answers to a certain item more similar to each other than to those of respondents interviewed by another interviewer.

Separation of design effects and interviewer effects is an issue that is addressed by only a few authors (Schnell and Kreuter, 2005). This is counter intuitive since whereas design effects arise in surveys with a cluster sample only, interviewer effects (in addition) are present in any face-to-face and even in a telephone survey – regardless of the sample design. Moreover, interviewer effects and design effects can interfere with each other. That is, homogeneity caused by one source of clustering (e.g. geographical) can have an influence on or be influenced by the other source (e.g. the interviewer). Ignoring either source of homogeneity can lead to biased or at least inefficient estimations.

The issue of separating design and interviewer effects and methods to determining the source of variation by use of a variance components model will be addressed in chapters 7.3 and 7.3 theoretically. An empirical investigation by means of a Monte Carlo simulation study is presented in Section 5.7.

2.4 Illustration and Motivation

The following illustration is based on results of selected runs of the Monte Carlo simulation study presented and discussed in Chapter 5. It shall demonstrate the inflation of variance of an estimator under a complex sample design (here: two-stage cluster sampling) compared to simple random sampling without replacement as described in the above example. This section serves a motivational purpose and does not explicitly clarify all elements used for illustration. This is done in the next chapter.

Figure 3 shows two overlaid histograms. Both histograms display the distribution

of values of the sample mean, $\bar{y} = \frac{1}{n}\sum_{i\in s}\sum_{j\in s_i} y_{ij}$, under two different sample designs. According to this formula, the mean is estimated in each of $I = 10\,000$ replicated draws from a population of size $N = 500\,000$ both under two-stage cluster sampling and simple random sampling without replacement (srswor). The population consists of $M = 1\,000$ clusters. Each cluster, in turn, consists of $N_i = 500, i = 1, \ldots, M$, ultimate sample units (for a detailed description of the universe(s) see Section 5.1).

The study variable, Y, is generated according to model M1 (2.9). In the present example σ is one and ρ is set to 0.10 which can be seen as a rather extreme value, chosen for illustrative purposes[2]. Table 2 gives summary statistics for the study variable in the population.

Table 2: *Summary statistics of the population*

	y
Min.	−4.4670
Q_{25}	−0.6745
Median	0.0002
Q_{75}	0.6744
Max.	4.9660
Mean	0.0000
SD	1.0002
Var	1.0004

In the Monte Carlo simulations, cluster samples and simple random samples each of size $n = 3\,000$ are repeatedly and independently drawn from this population. First, for the two-stage cluster sample design (clu2), in each sample $m = 150$ clusters are drawn on the first stage. At the second stage, $n_i = 20$ ultimate sample elements are drawn by srs from the sampled PSUs. Thus, every element in the sample has equal inclusion probabilities at both stages making weighting ignorable. Secondly, in each of the 10 000 simple random samples an identical number of $n = 3\,000$ ultimate sample elements is drawn from the population. Figure 3 shows the empirical distributions of sample means of these 10 000 replicated draws for the two sample designs in the same plot. The two histograms illustrate the different range of variation of the two distributions.

The red histogram – indicating the distribution of sample means based on the cluster sample design – has a larger variation in estimated sample means than the blue histogram. This can be seen from the thick tails of the red histogram and the lower density around the mean. The *empirical variance* of the 10 000 estimates under srswor is

$$Var\left(\bar{y}^{(\text{srswor})}\right) = \frac{1}{10\,000}\sum_{i=1}^{10\,000}\left(\bar{y}_i^{(\text{srswor})} - \bar{y}^{(\text{srswor})}\right)^2 = 3.2796 \cdot 10^{-4} \quad ,$$

2 Among this configuration also more and less extreme populations were generated with population values of $\rho = \{0.01, 0.02, 0.03, 0.04, 0.05, 0.10, 0.15, 0.20\}$ (see Chapter 5)

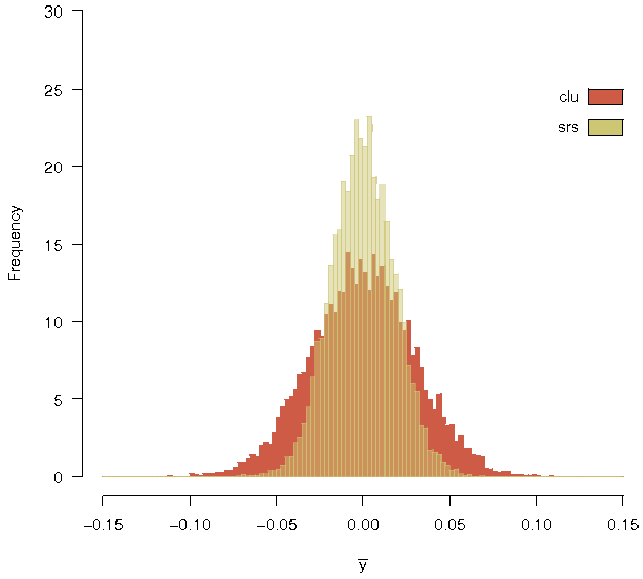


Figure 3: *Overlaid histograms of values of the estimator for the sample mean estimated under two-stage equal probability cluster sampling and simple random sampling without replacement from a clustered population with* $\rho = 0.10$

where $\bar{y}_{\bullet}^{(\text{srswor})}$ is the mean of estimated sample means over 10 000 iterations. The corresponding empirical variance of the distribution under cluster sampling is

$$Var\left(\bar{y}^{(\text{clu})}\right) = 8.7727 \cdot 10^{-4} \quad ,$$

which is

$$\frac{Var\left(\bar{y}^{(\text{clu})}\right)}{Var\left(\bar{y}^{(\text{srswor})}\right)} = \frac{8.7727 \cdot 10^{-4}}{3.2796 \cdot 10^{-4}} \approx 2.68$$

times the variance of the srswor estimate. This ratio is the *design effect*, *deff* which will be defined more thoroughly in the next chapter. Its square root, $\sqrt{deff} = deft$, is sometimes referred to as *design factor* (Kish, 1965,9). Table 3 gives some additional summary statistics for the estimates of simulated distributions of the sample mean under cluster and under simple random sampling without replacement.
The results in the table illustrate that the estimator of the sample mean of both sample designs is unbiased for the population mean. The distribution of sample means estimated on the basis of the srs replicates, however, shows less variability than the clu2 sample replicates.

Figure 4 consists of four smoothed scatterplots which illustrate the bivariate distribution of clu and srswor sample means for four different populations. The first population is generated such that hardly any homogeneity on the study variable within clusters appears which was achieved by a very small value of ρ (0.02) in the population. The second population was created assuming a slightly higher level

Table 3: *Summary statistics of the distribution of sample means of the study variable under two-stage equal probability cluster sampling and under srswor*

	$\bar{y}_{clu}$	$\bar{y}_{srswor}$
Min.	−0.1102	−0.0706
Q_{25}	−0.0200	−0.0122
Median	0.0001	0.0001
Q_{75}	0.0197	0.0123
Max.	0.1096	0.0679
Mean	0.0000	0.0000
SD	$2.9619 \cdot 10^{-2}$	$1.8110 \cdot 10^{-2}$
Var	$8.7727 \cdot 10^{-4}$	$3.2796 \cdot 10^{-4}$

of homogeneity within clusters ($\rho = 0.05$), the third ($\rho = 0.10$) was already described before, and the fourth shows a very high level of homogeneity within clusters ($\rho = 0.20$)[3]. Each plot of figure 4 shows the common distribution of 10 000 pairwise sample means under clu with an equal number of $\bar{b} = \frac{n}{m} = \frac{3000}{150} = 20$ elements per cluster and srswor of the same total sample size. Dark blue color indicates high density of observations in that region whereas a low density is indicated by light blue color. It can be seen that in the upper left plot the common distribution of the sample

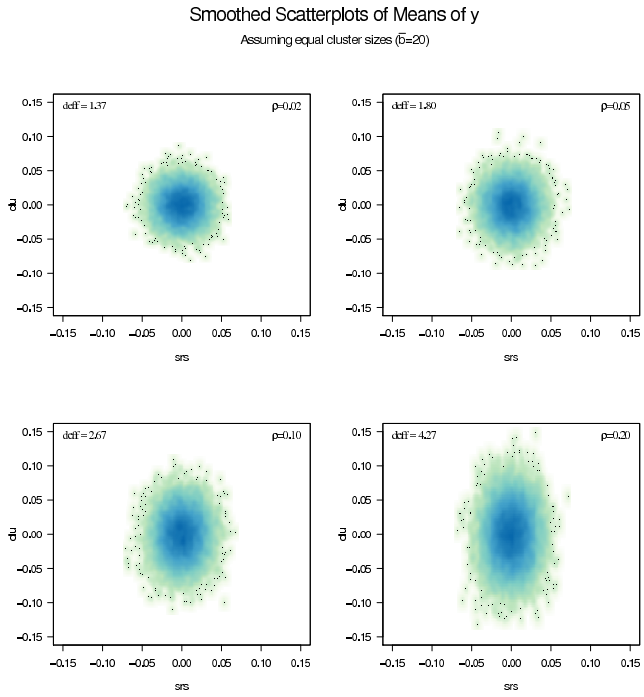
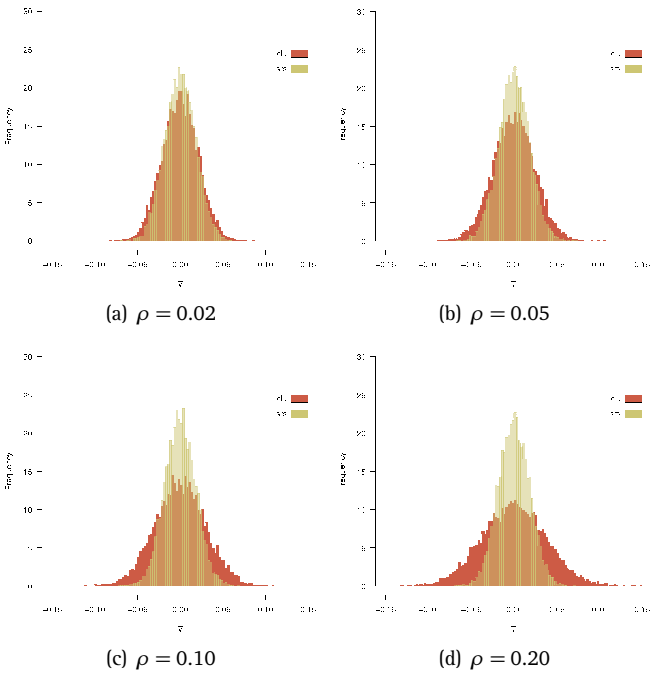


Figure 4: *Smoothed scatterplots based on a simulation of $I = 10\,000$ repeated draws under srswor and two-stage equal probability cluster sampling*

3 These populations were generated according to model M1 defined in (2.9).

means has almost the shape of a perfect circle. This indicates that the distributions of sample means under clu2 and under srs are very similar in respect to variation. The second plot on the upper right shows some deviation from a perfect circle – the points are scattered wider along the y-axis then along the x-axis. The third plot displays the distributions of sample means drawn from the clustered population with parameter $\rho = 0.10$ which was already analyzed graphically by the overlaid histograms in figure 3. Here the smoothed scatterplot has even more the shape of a football, indicating that variation in sample means is higher for cluster sampling than for srswor. The plot in the lower right corner finally shows the simulation results for sampling from the $\rho = 0.20$ population. Here the plot clearly shows the very imprecise nature of the sample mean estimated under cluster sampling as the plot is even more stretched along the y-axis.

Another way to look at this is, again, through overlaid histograms. Figure 5 shows the same distributions in four overlaid histograms. In each plot the area of intersection



(a) $\rho = 0.02$ (b) $\rho = 0.05$

(c) $\rho = 0.10$ (d) $\rho = 0.20$

Figure 5: *Overlaid Histograms based on a simulation of* 10 000 *repeated draws under srswor and two-stage equal probability cluster sampling (*$m = 150$, $\bar{b} = 20$*)*

of the two histograms is reciprocal to the magnitude of the design effect: large area of intersection indicates a low design effect (the variance of the estimate under cluster sampling is close to the variance under srswor) whereas less overlay indicates a large design effect. The evaluation of estimation strategies for design effects is at the focus of this thesis. The following chapter explicitly defines the concept of the design effect.

3 Design Effects

The examples in the previous chapter can be seen as motivatpoinal for a more thorough definition of the design effect laid out in this chapter. It opens with a discussion of design effects based on the seminal work by Kish (1965) in the following section. In Section 3.2, areas of application of design effects are illustrated and in the following section (3.3) an overview of the two approaches to the estimation of the design effect is given. Subsection 3.3.1 delineates the design-based estimation approach whereas the model-based approach is discussed in Subsection 3.3.2.

3.1 A Classical View on Design Effects

Design effects arise from a variety of divergences in real-world sample surveys from the (hypothetical) ideal of simple random sampling. Most prominent and intuitively appealing is the inflation of variance of an estimator due to clustering as highlighted in the illustration of the previous chapter. As already mentioned, it arises due to the fact that respondents living in the same geographical area are socialised in similar ways. Thus, their responses to survey items resemble each other more than they resemble the responses of respondents who live in another geographical area. However, the fact that the responses are more similar implies that, in terms of precision, the cluster sample data correspond to simple random sample data with less responses. This, in turn, means that the variance of an estimator (e.g. the HT-estimator) may be underestimated by the naive formula given by $\widehat{Var}\left(\hat{\bar{y}}\right) = \frac{s^2}{n}\left(1 - \frac{n}{N}\right)$. However, complex sample designs, and cluster sampling in particular, are not less precise than srs by definition as the following example shows.

Let us assume a population of elements grouped into M PSUs each of size N_i, $i = 1, \ldots, M$ and let $N = \sum N_i$. Furthermore, let $\bar{B} = \frac{N}{M}$ be the average cluster size. For the time being, let us assume that the PSUs are of equal size, so $N_i = B$. Finally, let y_{ij} denote the value of the variable of interest for the jth respondent in the ith cluster as before. Consequently, $y_i = \sum_{j=1}^{N_i} y_{ij}$ denotes the sum of study variable in the ith cluster. Again, the study variable follows the model M1 (2.9). A simple random sample of m clusters is drawn at the first stage and then all B elements of a a PSU are selected. The homogeneity of y introduced by geographical clustering leads to the design effect, *deff*, which is defined by Kish (1965, 162) as

$$deff_{\text{Kish}} = \frac{s_m^2/m}{s^2/n} \quad , \tag{3.1}$$

with $s_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(\bar{y}_i - \bar{y}\right)^2$ and $s^2 = \frac{1}{MB-1}\sum_{i=1}^{m}\sum_{j=1}^{B}\left(y_{ij} - \bar{y}\right)^2$ but shall here, following Lohr (1999, 239), be expressed more generally as

$$deff = \frac{Var_c\left(\hat{\theta}\right)}{Var_{srs}\left(\hat{\theta}\right)} \quad , \tag{3.2}$$

where $Var_c\left(\hat{\theta}\right)$ is the variance of the estimator $\hat{\theta}$ under the actual complex design (here: one-stage cluster sample design) and $Var_{srs}\left(\hat{\theta}\right)$ is the variance of the same estimator under a (hypothetical) simple random sample[4]. Put less formally, the design effect is the factor by which the variance of an estimator under a complex design is under or overestimated by the naive formula. The ratio

$$n_{eff} = \frac{n}{deff} \tag{3.3}$$

is referred to as the *effective sample size* and is the number of ultimate sample elements required in a srs which yields the same precision on a certain estimator as under a given complex sample design. Kish (1965, 162) showed that (3.1) can be expressed as

$$deff_{\text{one-stage}} = 1 + (B-1)\rho \quad , \tag{3.4}$$

if all B elements of a selected cluster are selected (*one-stage* or *cluster sampling*) and

$$deff_{\text{two-stage}} = 1 + (b-1)\rho \quad , \tag{3.5}$$

if b elements of a selected cluster are subsampled randomly (*two-stage sampling*). In expressions (3.4) and (3.5), ρ is the intraclass correlation coefficient which is defined as

$$\rho = \frac{\sum_{i=1}^{M}\sum_{j=1}^{B}\sum_{\substack{k=1\\k\neq j}}^{B}\left(Y_{ij} - \bar{Y}_U\right)\left(Y_{jk} - \bar{Y}_U\right)}{(B-1)(MB-1)S_Y^2} \tag{3.6}$$

with $S_y^2 = \frac{SST}{MB-1}$ with $SST = \sum_{i=1}^{M}\sum_{j=1}^{B}(Y_{ij} - \bar{Y}_U)^2$. This can also be written as

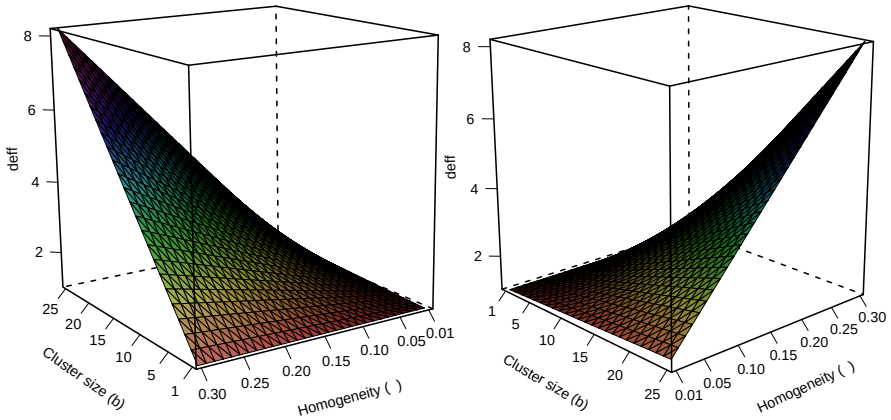
$$\rho = 1 - \frac{B}{B-1}\frac{SSW}{SST} \quad , \tag{3.7}$$

where $SSW = \sum_{i=1}^{M}\sum_{j=1}^{B}\left(Y_{ij} - \bar{Y}_{iU}\right)^2$. The domain of ρ ranges from $-\frac{B}{B-1}$ to one. The value of ρ can be negative when most or all of the total variation can be attributed to

4 The ratios of variances of the sample means in the above examples correspond to the design effect as defined by the above formula.

variation within clusters. This makes sense theoretically but will almost never occur practical applications.

It can easily be seen that the design effect depends on two parameters: If the cluster size, B or b, the second is the intraclass correlation coefficient, ρ. The following wireframe plot shows how *deff* changes as these two parameters vary. It can be seen



(a) 65° angle of view

(b) 245° angle of view

Figure 6: *deff as a function of b and* ρ

that when either of the parameters takes on a small value, an increase in the other one hardly increases the design effect. That is, big cluster sizes hardly matter as long as the intraclass correlation is small. Put the other way around, large intraclass correlation does not lead to a large design effect if the cluster size is small enough. Another way to look at this is through a level plot. In Figure 7 selected levels of *deff* are highlighted by solid lines. Any point on a line represents a combination of values of b and ρ that yield a design effect of the given magnitude. A design effect of 1.5, for example, can be achieved by a combination of $b = 11; \rho = 0.05$, $b = 6; \rho = 0.10$ or by $b = 3; \rho = 0.25$ as indicated by the grey crosses. The figure also illustrates that any two or more points on the plane with identical hue and saturation represent design effects of equal magnitude.

Let us now turn to the wide spread situation of unequal cluster sizes[5]. Kish (1965, 162) mentions the flexibility of expressions (3.4) and (3.5) in regard to their behaviour if cluster sizes vary. In such situations Kish (1989, 210) defines an estimator of the design effect given in (3.5) as

5 In the samples of participating countries of ESS, hardly any sample realizes a sample of PSUs of equal sizes but rather clusters of varying sizes as the selected sample designs discussed in Section 6.3 illustrate.

$$\widehat{deff} = 1 + (\bar{b} - 1)\rho \qquad . \tag{3.8}$$

In the more general case, when also ρ has to be estimated from sample data an appropriate estimator of the design effect is given by

$$\widehat{deff} = 1 + (\bar{b} - 1)\hat{\rho} \qquad , \tag{3.9}$$

where $\bar{b}$ is an estimate of the average cluster size and $\hat{\rho}$ is an adequate estimator of the intraclass correlation coefficient. Selected estimators of ρ are discussed in Chapter 4 which are evaluated in terms of bias and precision in Chapter 5.

In a scenario where weighting is necessary, additional variance is introduced in the HT estimator through variation of the design weights. A simple extension of the previous examples can illustrate such a situation: Instead of surveying an equal number of $n_i = B$ elements per selected PSU, let n_i vary. With a fixed cluster size in the population, $N_i = B$, inclusion probabilities of elements will vary between PSUs: $\pi_i = \frac{n_i}{N_i}$ and since $N_i = B$ we have $\pi_i = \frac{n_i}{B}$.

Kish (1987) proposed a formula "for determining the design effect in order to incorporate the effects due to both weighting needed to counter unequal selection probabilities and clustered selection". This formula is given by

$$\widehat{deff} = n \frac{\sum_{\ell=1}^{L} w_\ell^2 m_\ell}{\left(\sum_{\ell=1}^{L} w_\ell m_\ell\right)^2} \times \left[1 + (\bar{b} - 1)\rho\right] \qquad , \tag{3.10}$$

where m_ℓ denotes the number of observations in the ℓth weighting class, w_ℓ is the weight associated with the ℓth weighting class (not necessarily identical to PSUs).

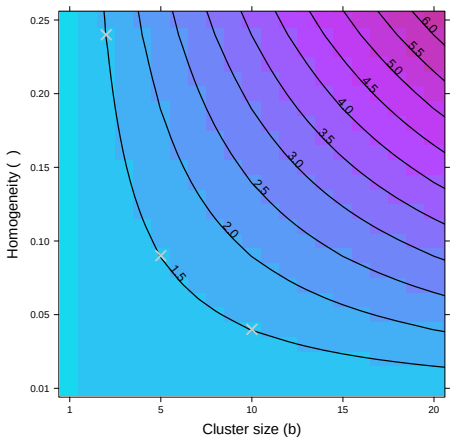


Figure 7: *Levels of deff by ρ and b*

Since variation in design weights stems from variation in inclusion probabilities, it is easy to see from (2.3) and (2.5) that unequal inclusion probabilities (i.e. π is such that at least one element is unequal to all other elements) lead to an increase in the variance and estimated variance of the HT estimator compared to the case of equal inclusion probabilities. The magnitude of increase is called the *design effect due to unequal inclusion probabilities*, $deff_p$, and is captured by the first term of (3.10). Hence, the estimator for the design effect of multi-stage cluster sampling with unequal inclusion probabilities can be written as

$$\widehat{deff} = \widehat{deff}_p \times \widehat{deff}_c \quad . \tag{3.11}$$

The concept of the design effect discussed above itself goes back to the ideas of several authors. Cornfield (1951) introduced a very similar concept which, in essence, he mainly needed to estimate a priori the required sample size of a cluster sample of census tracts. He defined the ratio of the standard error of the estimate of a certain prevalence rate under the assumption that the individual is the sampling unit to the standard error of an appropriate estimator of the prevalence rate estimator under the assumption of cluster sampling. This concept is to be seen as the inverse of the design effect as values above unity indicate a gain in efficiency when sampling Census tracts instead of individuals. Cornfield (1951) however uses the inverse of this ratio to define an early predecessor of the concept of the effective sample size which he calls "the factor by which the sample size, estimated by assuming that the individual is the sampling unit, must be increased if the tract is in fact the unit" (Cornfield, 1951, 660).

3.2 Areas of Application

Design effects basically serve two purposes: The first is to use design effects to estimate effective sample sizes at the planning phase of a survey. The second is to use them to correct naively estimated variance estimates and standard errors.

In the ESS, design effects are only used for the first purpose[6]. Since $deff_c$ varies from item to item a *typical design effect* due to clustering is estimated from data of the preceding round and the resulting value is taken as predictor for the upcoming round. Based on this value the minimum effective sample size of 1 500[7] is multiplied with the design effect in order to arrive at the required net sample size, i.e. the number of required interviews needed to fulfil the ESS specifications concerning the effective sample size. This means that the actual number of interviews will vary from country to country – sometimes quite significantly – depending on the sample designs and hence the predicted design effects.

Turning to the second purpose mentioned above, design effects should be used to correct standard errors, e.g. in classical hypothesis tests (Kish, 1989, pp. 209). Due to the fact that the variance of an estimator of a complex sample is underestimated

6 To the knowledge of the author, the ESS is the only social survey that incorporates design effects in this way.

7 In countries with less than two million inhabitants the minimum effective sample size is set to 800.

when calculated naively (i.e. assuming srs), the results of hypothesis tests will tend to be overoptimistic.

Unfortunately, only a subset of countries have given allowance to publish PSU labels along with the substantive data yet. This is why the ESS data archive publishes a typical design effect on-line for each country and round which can be used by the data analyst to correct standard errors. This policy ensures that substantive researchers are supplied at least with a good guess of *deff* if a country's privacy regulations do not permit publication of PSU identifiers. For all other countries, the researcher can estimate design effects for every item under study.

The aforementioned areas of application are based on different formulations of *deff*. Whereas, naturally, prediction of *deff* should be based on an adequate model, ex-post estimation of the design effect does not necessarily have to be model-based. In the following sections the model-based and the design-based approach to formulation and estimation of design effects is discussed in more detail.

3.3 Two Approaches to Design Effects

Although the definition of design effects is a purely design-based one, Gabler et al. (1999) have explicitly proposed a model-based justification for this formulation. Their model applies to a variety of real-word sample survey settings and is flexible enough to be adopted to even more complex settings, such as multiple design surveys (Gabler et al., 2006).

Skinner (1986) discussed design-based and model-based interpretations of the design effect thoroughly. In the ESS a model-based approach is relied upon to estimate the design effect. A detailed discussion and comparison of randomization-based and model-based design effects can be found in Gabler et al. (2010). In this section, a brief review of the most common methods for the estimation of design effects is given. Design-based as well as model-based approaches are discussed in the following two subsections.

3.3.1 Design-based Approach to Design Effects

Design-based estimation of the design effect is essentially based upon design-based methods of variance estimation. There exist numerous such variance estimation methods (Canty and Davison, 1999; Kish, 1989; Lohr, 1999; Münnich and Rässler, 2004; Münnich, 2004; Rao and Wu, 1988; Rao and Shao, 1999; Wolter, 1985). Various of them have been evaluated extensively in the DACSEIS project (Münnich, 2003a). A brief overview of the most important ones can be found in Canty and Davison (1999) and in Davison and Sardy (2007). In the following, the two most commonly used methods – Taylor Linearisation and Jackknife Repeated Replication (JRR) – are discussed.

For design-based estimation of the design effect mainly JRR and Balanced Repeated Replication (BRR) have been used (Gonzalez and Foy, 2004; Lago et al., 1987a). These techniques are replication based methods and as such rely on a large number

of repeated draws from sample data. Thus, they tend to be computationally intensive. Taylor Linearisation, in contrast, is an analytical approximative technique.

Looking at estimation of design effects from a design-based point of view, methods are required that efficiently and unbiased estimate equation (3.2): $\widehat{deff}_c = \frac{\widehat{Var}_c(\hat{\theta})}{\widehat{Var}_{srs}(\hat{\theta})}$. The main problem is to find good estimates for the numerator of this ratio. One of the problems of the design-based approach to estimation of the design effect is, however, to find an adequate estimator for $\widehat{Var}_{srs}(\hat{\theta})$. A commonly used solution is to treat the data of the cluster sample naively as having arisen from a simple random sample with replacement and estimate the variance of $\hat{\theta}$ as $\widehat{Var}_{srs}(\hat{\theta})$ by $\frac{s^2}{n}$. The simulation study on the Monte Carlo estimation of the true design effect in Section 5.3, however, uses a separate srs of the same (expected) sample size like the corresponding cluster sample (for details see Section 5.2.1). This strategy serves as a baseline in the following sections which evaluate different design- and model-based estimators of the design effect. In addition, a small Monte Carlo study in Section 5.4.1.4 evaluates the effects of using the naive formula for the variance estimator of the sample mean on data from a cluster sample.

3.3.1.1 Taylor Linearisation

Linearisation methods in variance estimation are based on a theorem by Taylor (1969). Let there be K estimators of the population total, $\tau_1, \ldots, \tau_K$ and let g be a function on these estimators. Then an approximate estimator of the variance of an estimator, $\hat{\theta}$, is, according to Münnich (2008, 323), given by

$$\widehat{Var}\left(\hat{\theta}\right) = \text{var}\left(\sum_{k=1}^{K} \frac{\partial g\left(\hat{\tau}_1, \ldots, \hat{\tau}_K\right)}{\partial \hat{\tau}_k}\left(\hat{\tau}_k - \tau_k\right)\right) \quad . \tag{3.12}$$

The linearisation method can be used in cluster sampling to estimate the variance of the combined ratio estimator of m cluster totals to m weighted cluster sizes: $r = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}y_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}}$. This estimator is used as an estimator for the population mean. The Taylor linearisation estimator of the variance of this estimator is given by

$$\widehat{Var}_c^{Taylor}(\hat{\theta}) = \frac{\sum_{i=1}^{m} \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}}{\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\right)-1}\left(z_i^2 - \frac{z^2}{\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}}\right)}{\left(\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}\right)^2}, \tag{3.13}$$

where $z_i = \sum_{j=1}^{n_i} w_{ij} y_{ij} - \left(r \cdot \sum_{j=1}^{n_i} w_{ij} \right)$. For a general discussion of linearised variance estimators of ratios see also Binder (1996) and Demnati and Rao (2002). The design-based estimator of the design effect for the sample mean is then given by

$$\widehat{deff}^{Taylor} = \frac{\widehat{Var}_c^{Taylor}(\hat{\theta})}{\widehat{Var}_{srs}(\hat{\theta})} \quad . \tag{3.14}$$

The above estimator is easily adopted to the case when y has dichotomous outcome since y_{ij} can then be regarded as an indicator variable taking the value of one if the jth person in the ith cluster has a positive outcome on the the study variable and zero otherwise.

Taylor Linearization makes most sense when the variance estimator is has non-linear terms which makes it difficult to compute directly. Within the scope of the simulation studies in Chapter 5 the estimator in 3.14 is evaluated as one candidate of the class of design-based estimators for *deff* for the sample mean and for the median (see Section 5.4.2).

3.3.1.2 The Jackknife Method

The Jackknife method – also referred to as Jackknife Repeated Replication – is described in detail in Wolter (1985, chapter 4). A brief discussion of its merits and shortcomings is given in Münnich (2008, pp. 325). As applied to variance estimation in cluster sampling, the basic approach is to run as many iterations as there are PSUs. In each run, however, the ith PSU is omitted for the calculation of the variance. The average of the resulting C variance estimations is then taken as the final variance estimation. Stated formally, the JRR estimator is defined as in (Gonzalez and Foy, 2004, 82)

$$\widehat{Var}_c^{JRR}(\hat{\theta}) = \frac{1}{m} \sum_{i=1}^{m} \left(\hat{\theta}(J_i) - \hat{\theta}(S) \right)^2 \quad , \tag{3.15}$$

where $\hat{\theta}(S)$ refers to the estimator of parameter θ based on data of the whole sample and $\hat{\theta}(J_i)$ is the estimator of the same parameter based on the ith JRR replicate sample, i.e. excluding the ith PSU for each estimation.

The JRR design-based design effect, then, is given by

$$\widehat{deff}^{JRR} = \frac{\widehat{Var}_c^{JRR}(\hat{\theta})}{\widehat{Var}_{srs}(\hat{\theta})} \quad . \tag{3.16}$$

The flexibility of the JRR method comes into play when the variance estimator of a parameter under complex sample designs has no closed form. In the case of non-parametric point estimators like the median, the JRR estimator follows the same principle as in the case of parametric estimators like the HT estimator of the population

total or the population mean[8]. The JRR estimator is adopted for binary variables by substituting an appropriate estimator of the overall success rate, $\hat{\pi}$, for $\hat{\theta}$ and its variance estimator under srs for $\widehat{Var}_{srs}(\hat{\theta})$.

3.3.2 Model-based Approach to Design Effects

Model-based estimation differs from the design-based approach mainly in the assumptions about the data generating process and hence the way estimators of population parameters have to be thought of. Where the design-based approach focuses on the set of possible samples that could have been drawn from a population in order to protect itself against model failure, the model-based perspective conditions inferences on the realized sample and is thus commonly judged more robust – iff the model holds.

In many cases, however, following a design-based or a model-based approach reveals identical formulae and, hence, identical numerical values for a number of common estimators. This is also true for the estimation of the design effect. A model-based version of the design effect has been suggested by Gabler et al. (1999) and further developed by Gabler et al. (2006) and Lynn and Gabler (2005).

The model-based approach is generally attractive due to its flexibility to deviations in the data from a suggested model. In an analogous manner, the model-based estimator of the design effect presented in the following can be thought of as an instance of a wider class of estimators based on models that account for even more complex situations.

A model-based estimator of the design effect assumes a model from which the data are generated. The following delineation describes such a model and explains its interrelation with the proposed model-based estimator of *deff*. The following discussion is closely lent on Gabler et al. (1999) and Gabler et al. (2006).

Let $q_{i\ell}$ be the number of observations in the ith cluster and the ℓth weighting class, let $q_\ell = \sum_{i=1}^{m} q_{i\ell}$ be the number of observations in the ℓth weighting class, and let $n = \sum_{\ell=1}^{L} q_\ell$, as previously, denote the sample size. Further let $n_i = \sum_{\ell=1}^{L} q_{i\ell}$ be the number of observations in the ith cluster. Hence, $\bar{b} = \frac{1}{m}\sum_{i=1}^{m} n_i = \frac{n}{m}$ is the average cluster size.

8 However, with a non-smooth statistik like the median, JRR may have problems with consistency.

Taking into account the usual design-based HT estimator of the population mean, $\bar{y}_w = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij} y_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}}$, Gabler et al. (1999) assume the following model (M1):

$$Var(y_{ij}) = \sigma^2 \text{ for } i = 1,\ldots,m; j = 1,\ldots,n_i \tag{3.17}$$

$$Cov(y_{ij}, y_{i'j'}) = \begin{cases} \rho\sigma^2 & \text{if } i = i'; j \neq j' \\ 0 & \text{otherwise.} \end{cases} \tag{3.18}$$

A second model (M2) specifies the distribution of the y_{ij} in the following way:

$$Var(y_{ij}) = \sigma^2 \text{ for } i = 1,\ldots,m; j = 1,\ldots,n_i \tag{3.19}$$

$$Cov(y_{ij}, y_{i'j'}) = 0 \text{ for all } (i,j) \neq (i',j'). \tag{3.20}$$

Let $Var_{M1}(\bar{y}_w)$ be the variance of the weighted sample mean under model M1 and let $Var_{M2}(\bar{y})$ be the variance of the overall sample mean, $\bar{y} = \sum_{i=1}^{m}\sum_{j=1}^{n_i} \frac{y_{ij}}{n}$, under M2, respectively. Under M2, the variance of $\bar{y}$, however, turns out to be given by $Var_{M2}(\bar{y}) = \frac{\sigma^2}{n}$. Then the design effect is defined as

$$deff = \frac{Var_{M1}(\bar{y}_w)}{Var_{M2}(\bar{y})} \quad . \tag{3.21}$$

According to Gabler et al. (1999) *deff* can be expressed as

$$deff = n \frac{\sum_{\ell=1}^{L} w_\ell^2 q_\ell}{\left(\sum_{\ell=1}^{L} w_\ell q_\ell\right)^2} \times \left[1 + (b^* - 1)\rho\right] \quad , \tag{3.22}$$

where

$$b^* = \frac{\sum_{i=1}^{m}\left(\sum_{j=1}^{n_i} w_{ij}\right)^2}{\sum_{i=1}^{m}\sum_{j=1}^{n_i} w_{ij}^2} \quad . \tag{3.23}$$

The quantity ρ servers as a measure of homogeneity and as such gives information about the similarity of the clustered elements. High values of ρ imply very similar

values on the variable under study within the clusters and very dissimilar values between clusters. The usual ANOVA estimator of ρ is given by

$$\hat{\rho}^{(\mathrm{AOV})} = \frac{MSB - MSW}{MSB + (K-1)MSW} \quad , \tag{3.24}$$

where

$$MSB = \frac{SSB}{m-1}$$

with $SSB = \sum_{i=1}^{m} n_i \left(\bar{y}_i - \bar{y}\right)^2$ and

$$MSW = \frac{SSW}{n-m}$$

with $SSW = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$ and

$$K = \frac{1}{(m-1)} \left(n - \sum_{i=1}^{m} \frac{n_i^2}{n} \right).$$

The above model has the advantage that it applies in many real-world situations. In the ESS, for example, the model-based design effect is estimated according to the above formula in countries where sampling was done either using a) unequal inclusion probabilities, b) clustering or c) both. What makes it even more useful is that it can also be applied to *multiple design surveys*. Gabler et al. (2006) showed that (3.22) has a generalized form that allows to calculate a weighted average of *deff* over multiple domains in a sample. A *domain* is defined by a subset of the sample in which the design effect is one, e.g. a part of the sample where instead of a complex sample design, a srs of ultimate sample units is drawn. A multiple sample design with two domains, for example, is realized in Poland where in large cities a srs of respondents and in rural areas a cluster sample is drawn (for details see Section 6.3.3). Here, the design effect is predicted only for the clustered part of the sample and combined with the design effect of the srs part of the sample according to the proposed model.

4 Measures of Homogeneity

Homogeneity refers to the degree by which respondents belonging to the same predefined structural entity, or cluster, resemble each other. This entity may be defined as a geographical region in the most common case when considering the design effect due to clustering or the trait of being interviewed by the same interviewer (Gabler and Lahiri, 2009). In both settings respondents tend to be more similar to other respondents belonging to the same structural entity than to respondents of another structural entity. This implies that large differences on an item will mainly be found at respondents belonging to different clusters whereas the odds of finding large differences between respondents of the same cluster are rather small. However, instead of the term similarity the more stringently defined term of *homogeneity* shall be used. Homogeneity refers to the relative similarity of elements within the same cluster to their expected similarity with all other elements. Quantification of the degree of homogeneity within the scope of this thesis is by means of the intraclass correlation coefficient (ICC), denoted by ρ and its estimators, $\hat{\rho}$.

This chapter first briefly reviews the existing literature on estimation methods for ρ. Then, in Section 4.2, estimators for continuous and in Section 4.3 special estimators for binary data are presented.

4.1 Overview of Estimation Methods

Estimation and interpretation of the intraclass correlation coefficient is a heavily debated topic. A very early paper focusing on a specific topic is the one by Irwin (1946) who discusses ways of interpreting negative values of ICC. The effects on intraclass correlation on confidence intervals and the results of significance tests were already discussed by Walsh (1947). A non-parametric estimator if the intraclass correlation coefficient was proposed by Rothery (1979) which employs the probability of concordance values. Clemmer and Kalsbeek (1984) propose an alternative "proportion variation method" to for the estimation of ρ. Donner (1986) gives a nice overview of inferential methods for the evaluation of point and interval estimates of ρ. Choi (1987) discusses a closed-form direct estimator of the intraclass correlation, going back to Cohen (1960) and Kleinman (1973) with which the variances of the correlation estimators shall be reduced. An estimator and its variance estimator of the intraclass correlation for categorical data is given in Choi (1989). The estimator underlies a one-way ANOVA model within each level of the variable. As computational power grew rapidly in the 90's, iterative methods came into the focus of researchers. Paul (1990) for example presents an Maximum-Likelihood estimator and its variance estimator based on estimating equations which can only be solved iteratively, for example by the ZBRENT subroutine of IMSL (Paul, 1990, 553). McGraw and Wong (1996) present a comprehensive overview of different types of estimators applicable in one- and two-way random and mixed models both with and without interaction effects. As the proficiency of the authors is in psychology, the paper cover ρ as the rate of homogeneity as a special case.

An evaluation of the use of ρ as an analytical tool in a small cluster sample survey[9] is given in Fields (1970). Measures of intraclass correlation can also be used to assess inter-rater reliability. Shrout and Fleiss (1979) give an example of how to choose among different estimator of ρ depending on the data generating model which is assumed. An illustration of the influence of cluster sizes on a specific estimator of ρ and also the design effect is given in Thomas et al. (1983) by artificially generating clusters from Enumeration Districts (ED) of the 1980 US Census. The paper by Martinez and Brogan (1984) empirically evaluates the behaviour of two different types of estimators of ρ, namely the *design effect method*, $\rho = \frac{\widehat{deff} - 1}{\bar{b} - 1}$, and a classical ANOVA estimator. Mak (1988) discusses the asymptotic variance of the variance component based estimator of ρ for binary data and presents an evaluation study based on littermate data. An empirical investigation of the homogeneity of smoking behaviour of students within classes and the effect of a loss in precision is given in Siddiqui et al. (1996). An overview of a large number of estimators of the intraclass correlation coefficient for binary data and an evaluation based on a simulation study is given in Ridout et al. (1999). In fact, many of the estimators presented herein are also evaluated in the Monte Carlo simulations presented in Chapter 5. Paul et al. (2003) give an overview of extensive simulation studies of estimators for dichotomous variables. They find that a version of the estimator based on Quadratic Estimation Equations is very precise and has very little bias. However, the usual ANOVA estimator also performs very well (Paul et al., 2003, 518). Rodríguez and Elo (2003) discuss the use of random effects models to estimate intraclass correlation for binary variables using probit, logit, and complementary log-log models. Zou and Donner (2004) propose a method for estimation of confidence intervals of intraclass correlation estimators for binary data.

Due to the large amount of estimators proposed, it is essential to define some structure that helps distinguish between different types of estimators. On the one hand, estimators can be distinguished by the data scale type (continuous versus binary) for which they are appropriate. Certain estimators, for example, are defined for binary data only and do not work for continuous data (and the other way around). Section 4.2 describes estimators for continuous data and Section 4.3 is focused on estimators for dichotomous items.

A second line of differentiation concerns the definition of estimators themselves. Do estimators directly use the (estimated) correlation structure of the data, are they based on an ANOVA decomposition, or do they rely on the variance components of a random effects model? Estimators belonging to the first two classes will be grouped together and named *classical* estimators whereas estimators of the third class will be referred to as random-effects based estimators. Estimators for continuous data belonging to the first group are discussed in Section 4.2.1 and random-effect model-based estimators in Section 4.2.2. The same differentiation is made also for estimators for binary data: Section 4.3.1 presents classical estimators whereas Section 4.3.4 fo-

9 The 1968 Michigan Detroit Area Study, (Fields, 1970, pp. 594).

cuses on the definition of estimators based on random effect models for dichotomous data. These kinds of models (Generalized Linear Models, GLM) are, in fact, worth more detailed consideration. An introduction into the basic model and estimation techniques is given in Chapters 7.3 and 7.3 which give more insight into the specific problems associated with the derived estimators of ρ.

4.2 Estimators for continuous data

Among the rather classical estimators like the ANOVA estimator and a set of F-statistic based estimators (see 4.2.1), this section also describes estimators based on variance components of random effect models. For the time being, the reader has to be referred to Section 4.2.2 and Chapter 7.3 for a detailed discussion of the specification and estimation of the model. Generally, we shall distinguish between the following estimators for continuous data:

1. Estimators based on ANOVA model and F-statistics:
   $\hat{\rho}^{(\mathrm{AOV})}$ The classical ANOVA estimator
   $\hat{\rho}^{(\mathrm{F})}$ An estimator based on F-statistics
   $\hat{\rho}^{(\mathrm{F2})}$ Similar to the one above but denominator also based on F-statistic
   $\hat{\rho}^{(\mathrm{FR})}$ Estimator as implemented in the Hmisc package of R
2. Estimators based on a variance decomposition of a random effects model. These estimators differ in respect to the estimation method used for the random effects model. Estimation of a random effects model can be based on a number of different techniques of which the effects on the quality of estimators of the intraclass correlation coefficient of the following three methods will be investigated:
   $\hat{\rho}^{(\mathrm{ML})}$ Maximum Likelihood optimization
   $\hat{\rho}^{(\mathrm{REML})}$ Restricted Maximum Likelihood optimization
   $\hat{\rho}^{(\mathrm{Laplace})}$ Laplace approximation assuming Gaussian family and identity link

A detailed description of these estimators is given in the following subsections.

4.2.1 ANOVA Estimator and F-statistic based Estimators

The most intuitive and appealing of all classical estimators is the ANOVA estimator, $\hat{\rho}^{(\mathrm{AOV})}$, which is based on an ANOVA of the study variable with clusters as grouping factors. It is defined as in (3.24) on page 45.

There exist another class of estimators, namely those based on F-statistics. One of those is $\hat{\rho}^{(\mathrm{F})}$, the basic F-statistic estimator. For centered data, i.e. $\bar{y} = 0$, it is given by

$$\hat{\rho}^{(\mathrm{F})} = \frac{\sum_{i=1}^{m} \sum_{\substack{j=1 \\ j' \neq j}}^{n_i} \frac{y_{ij} y_{ij'}}{n_i (n_i - 1)}}{Var(y) m} \quad . \tag{4.1}$$

A modified version is given by

$$\hat{\rho}^{(\mathrm{F2})} = \frac{F-1}{F-1+\left(\frac{n}{m}\right)} \quad , \tag{4.2}$$

where $F = \frac{MSB}{MSW}$. For equal cluster sizes $\hat{\rho}^{(\mathrm{AOV})}$ and $\hat{\rho}^{(\mathrm{F2})}$ are equal.

The F-statistic estimator as implemented in the Hmisc package of R is given by

$$\hat{\rho}^{(\mathrm{FR})} = \frac{\left[R\frac{n-m}{(1-R)m}-1\right]\bar{b}}{1+\left[R\frac{n-m}{(1-R)m}-1\right]\bar{b}} \quad , \tag{4.3}$$

where

$$R = 1 - \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} n_i^{-1}\left(y_{ij}-\bar{y}_i\right)^2}{\sum_{i=1}^{m}\sum_{j=1}^{n_i} n_i^{-1}\left(y_{ij}-\bar{y}\right)^2} \quad .$$

4.2.2 Estimators based on Random Effects Models

Assume the study variable is generated according to the following one-way ANOVA model:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \quad , \tag{4.4}$$

where μ is the mean, the α are random effects (due to cluster membership) and ϵ are independent error terms as well as independent from α. Both α and ϵ are normally distributed with mean zero but α has variance σ_α^2 and ϵ has variance σ_ϵ^2. According to McGraw and Wong (1996, 35), the intraclass correlation coefficient of the study variable under the above model is generally defined by

$$\rho^{(\mathrm{RE}^*)} = \frac{\sigma_\alpha^2}{\sigma_\alpha^2+\sigma_\epsilon^2} \quad . \tag{4.5}$$

The star in the superscript indicates that the formula does not directly represent an estimator per se but rather a class of estimators. Specific estimators of the RE class of $\hat{\rho}$ differ in respect of the methods which are used to estimate the variances in the equation above. The quality (i.e. bias and precision) of $\hat{\rho}^{(\mathrm{RE}*)}$ directly depends on the quality of these estimation methods. Among these estimation methods for random effect models, mainly Maximum Likelihood (ML), Restricted Maximum Likelihood (REML) and Laplace Approximation methods are discussed in the literature. ML and REML methods are discussed in Harville (1977) and their implementation in R is illustrated in Faraway (2006, pp. 153). A comprehensive summary and a discussion of the advantages and disadvantages of these three estimation methods and their implications for the estimation of ρ is given in Section 7.3.

The estimator of ρ based on an ANOVA decomposition of the variances in the above model can be written explicitly as

$$\hat{\rho}^{(\mathrm{RE-AOV})} = \frac{\hat{\sigma}^2_{\alpha,(AOV)}}{\hat{\sigma}^2_{\alpha,(AOV)} + \hat{\sigma}^2_{\epsilon,(AOV)}} = \frac{\frac{MSB - MSW}{n}}{\frac{MSB - MSW}{n} + MSW},$$

where MSB and MSW are the usual mean squared errors (within and between groups) of the ANOVA decomposition. Estimators of ρ belonging to the RE* class which are based on variance components deduced by alternative estimation techniques (e.g. ML, REML, and Laplace) are referred to as $\hat{\rho}^{(\mathrm{RE-ML})}$, $\hat{\rho}^{(\mathrm{RE-REML})}$, and $\hat{\rho}^{(\mathrm{RE-Laplace})}$ or, for short $\hat{\rho}^{(\mathrm{ML})}$, $\hat{\rho}^{(\mathrm{REML})}$, and $\hat{\rho}^{(\mathrm{Laplace})}$.

Estimators of the RE class, besides $\hat{\rho}^{(\mathrm{RE-AOV})}$ above, cannot be negative by definition as equation (4.5) suggests. This can be seen as a problem if they are used in the formula for $\widehat{deff}_c$, since ρ in (3.6) can take on negative values. As mentioned earlier, this is the case when a great share of the total variance can be explained by within cluster variation. In that case, thinking in terms of efficiency, cluster sampling would be even more efficient than simple random sampling since information gained by surveying all selected respondents in each cluster contributes to a large amount to total variance. In the another extreme scenario, all elements in clusters resemble each other perfectly. Thus, after having gathered information on one respondent of a cluster, there is no gain in surveying any further person since he does not contribute any new information (and hence no additional variation) to the study variable.

4.3 Dichotomous Variables

A large number of estimators of the intraclass correlation coefficient for binary variables has been proposed and a number of evaluation studies have been undertaken to investigate the behaviour of some of these estimators (see Section 4.1). Among these studies, the paper by Ridout et al. (1999) gives a very good overview of a wide range of classical estimators. Paul et al. (2003) considers 26 estimators based on advanced estimation methods (e.g. Maximum Extended Beta-Binomial Likelihood).

The estimators of ρ for binary data presented in the following can be classified in the same way as in the preceding section. On the highest level we shall distinguish between so called classical estimators (i.e. based on ANVOA models or F-statistic based estimators) and estimators based on variance components of random effects models. Most of the classical estimators are taken from the overview paper by Ridout et al. (1999) and are notationally adopted.

To account for the special case of binary data, we have to slightly adopt notation. Let, without loss of generality, y_{ij} be one if the jth element of the ith cluster has a positive outcome on the study variable (e.g. owns a TV, took part in the last national

vote, etc.). Further, let $y_i = \sum_{j=1}^{n_i} y_{ij}$ be the sum of positive outcomes within the ith cluster.

4.3.1 Classical ANOVA Estimator

As well as when faced with continuous data, the most intuitive estimator for binary data can be seen to be the ANOVA estimator. Also for binary data $\hat{\rho}^{(\mathrm{AOV})}$ relies on the mean square error between and within groups (i.e. clusters). These quantities, however, have to be estimated differently with dichotomous items. Apart from that, the basic logic behind the AOV estimator stays the same. Hence, also for binary items the estimator of the intraclass correlation coefficient is defined as

$$\hat{\rho}^{(\mathrm{AOV})} = \frac{MSB - MSW}{MSB + (K-1)MSW} \quad , \tag{4.6}$$

where $K = \frac{1}{(m-1)}\left(n - \sum_{i=1}^{m} \frac{n_i^2}{n}\right)$ and the quantities MSB and MSW being the appropriate mean square error between and mean square error within clusters. The AOV estimator has the same merits and drawbacks as in the case of continuous data (see section 4.2.1).

4.3.2 Estimators based on Moments

With binary data a class of estimators based on the first moment was defined by Kleinman (1973). Put most generally, estimators of this class are defined as

$$\hat{\rho}^{(\mathrm{K}^*)} = \frac{SSW_{(w)} - \tilde{\pi}_w\left(1-\tilde{\pi}_w\right)\sum_{i=1}^{m}\frac{w_i\left(1-w_i\right)}{n_i}}{\tilde{\pi}_w\left(1-\tilde{\pi}_w\right)\left[\sum_{i=1}^{m} w_i\left(1-w_i\right) - \sum_{i=1}^{m}\frac{w_i\left(1-w_i\right)}{n_i}\right]} \quad , \tag{4.7}$$

with the weighted sums of squares within clusters, $SSW_{(w)} = \sum_{i=1}^{m} w_i\left(\tilde{\pi}_i - \tilde{\pi}_w\right)^2$, $\tilde{\pi}_w = \sum_{i=1}^{m} w_i\tilde{\pi}_i$ and $\tilde{\pi}_i = \frac{y_i}{n_i}$. The weights, w_i, defined in (4.8) and (4.9), must be scaled to sum to one (Ridout et al., 1999, 138). Different estimators of this class differ in respect with the choice of the weights and hence the magnitude of the weighted overall proportion, $\tilde{\pi}_w$. Based on the above formula, Kleinman (1973) proposed the KEQ and KPR estimators assuming equal (KEQ) and unequal (KPR; i.e. proportional to cluster size) weights:

$$\hat{\rho}^{(\mathrm{KEQ})}: \quad w_i = \frac{1}{m} \tag{4.8}$$

$$\hat{\rho}^{(\mathrm{KPR})}: \quad w_i = \frac{n_i}{n} \quad . \tag{4.9}$$

Another estimator relying on moments[10] was proposed by Yamamoto and Yanagimoto (1992). They call it the "unbiased moment estimator" (Yamamoto and Yanagimoto, 1992, 274) as they state that the estimation equation it is derived from has expectation zero. The UBE estimator is defined as

$$\hat{\rho}^{(\mathrm{UBE})} = 1 - \frac{nK(m-1)MSW}{\sum_{i=1}^{m} y_i \left[K(m-1) - \sum_{i=1}^{m} y_i\right] + \sum_{i=1}^{m} y_i^2} \quad . \tag{4.10}$$

In addition to the $\hat{\rho}^{(\mathrm{UBE})}$ estimator, Yamamoto and Yanagimoto (1992) also suggest a "raw" (RAW) and a "central" (CEN) estimator. They take into account the fact that sometimes raw and central moment systems are used (Yamamoto and Yanagimoto, 1992, 275). A so called "stabilized moment estimator" (STA) by Tamura and Young (1987) is designed to overcome the bias of the MLE estimator. The STA estimator is identical to the CEN estimator with the exception that it incorporates a stabilizing term that was empirically derived by Tamura and Young (1987) in a simulation study. Due to the expected small gain in precision and only minor reduction in bias, these estimators will not be considered hereafter.

Mak (1988) assumes a beta-binomial distribution model (see for example Ridout et al. (1999, pp. 137)) for the study variable with parameters α and β denoting the probability that two elements show the same response given they belong to the same and different clusters, respectively. Based on the formulation in Fleiss and Cuzick (1979), Mak (1988, 139) proposes an unbiased estimator of ρ which is given by

$$\hat{\rho}^{(\mathrm{MAK})} = 1 - \frac{(m-1)\sum_{i=1}^{m} \frac{y_i(n_i - y_i)}{n_i(n_i-1)}}{\sum_{i=1}^{m} \frac{y_i^2}{n_i^2} + \left(\sum_{i=1}^{m} \frac{y_i}{n_i}\right)\left(m - 1 - \sum_{i=1}^{m} \frac{y_i}{n_i}\right)} \quad . \tag{4.11}$$

This estimator is based on an unbiased estimator of β which is the average of $\frac{m(m-1)}{2}$ cluster-wise estimates of β. The estimate of β for the pair of the ith and the qth cluster is given by $\beta_{(i,q)} = 1 - \frac{y_i(n_q - y_q) + (n_i - y_i)y_q}{n_i n_q}$.

The $\hat{\rho}^{(\mathrm{MAK})}$ estimator is closely leant on an estimator proposed by Fleiss and Cuzick (1979) who estimate β by $1 - 2\hat{\pi}(1-\hat{\pi})$ where $\hat{\pi} = \frac{\sum_{i=1}^{m} y_i}{\sum_{i=1}^{m} n_i}$ is the overall rate of success in the data. Their estimator is given by

$$\hat{\rho}^{(\mathrm{FLC})} = 1 - \frac{1}{(NM-k)\hat{\pi}(1-\hat{\pi})} \sum_{i=1}^{m} \frac{y_i(n_i - y_i)}{n_i} \quad , \tag{4.12}$$

10 In fact, they developed the estimator as an estimator for the shape parameter Φ in the general probability function of the beta-binomial distribution: $f(x; n, \mu, \Phi) = \frac{\binom{n}{x}\prod_{r=0}^{x-1}\left(\mu + r\frac{\Phi}{1-\Phi}\right)\prod_{r=0}^{n-x-1}\left(1-\mu+r\frac{\Phi}{1-\Phi}\right)}{\prod_{r=0}^{n-1}\left(1+r\frac{\Phi}{1-\Phi}\right)}$

where N and M denote the cluster size and the number of clusters in the population, respectively.

4.3.3 Direct Estimation of Correlation Structure

The most *direct* way to estimate the intraclass correlation coefficient is to directly estimate the correlations within groups (Donner, 1986). The resulting estimator of ρ, however, has the undesirable property that too much weight is assigned to large clusters. A natural approach is hence to assign weights to the group-wise correlations. A general form of such a weighted direct estimator, proposed by Williams (1982), is given by

$$\hat{\rho}^{(\mathrm{PW})} = \frac{\sum_{i=1}^{m} w_i \sum\sum_{j \neq l}^{n_i} (y_{ij} - \hat{\mu})(y_{il} - \hat{\mu})}{\sum_{i=1}^{m} w_i (n_i - 1) \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu})} \quad , \tag{4.13}$$

where $\hat{\mu} = \sum_{i=1}^{m} w_i (n_i - 1) \sum_{j=1}^{n_i} y_{ij}$ is the weighted mean for a given study variable Y and weights scaled to satisfy $\sum_{i=1}^{m} n_i (n_i - 1) w_i = 1$. In the case of binary data (4.13) reduces to (see Ridout et al. (1999, 139))

$$\hat{\rho}^{(\mathrm{PWB})} = \frac{\sum_{i=1}^{m} w_i y_i (y_i - 1) - \hat{\mu}^2}{\hat{\mu}(1 - \hat{\mu})} \quad . \tag{4.14}$$

Different choices of weights yield different estimators of $\hat{\mu}$ and hence define different estimators of the $\hat{\rho}^{(\mathrm{PWB})}$ class. In the most simple case weights are equal and thus equal weight is attributed to each pair of observations. This way the estimator for μ is defined as

$$\hat{\mu}_{(\mathrm{PEQ})} = \frac{\sum_{i=1}^{m} (n_i - 1) y_i}{\sum_{i=1}^{m} (n_i - 1) n_i} \tag{4.15}$$

and the PEQ estimator for ρ is

$$\hat{\rho}^{(\mathrm{PEQ})} = \frac{1}{\hat{\mu}_{(\mathrm{PEQ})} \left(1 - \hat{\mu}_{(\mathrm{PEQ})}\right)} \left[\frac{\sum_{i=1}^{m} y_i (y_i - 1)}{\sum_{i=1}^{m} n_i (n_i - 1)} - \hat{\mu}^2_{PEQ} \right] \quad . \tag{4.16}$$

When weights for each group are equal, $w_i = \dfrac{1}{m n_i (n_i - 1)}$, we have

$$\hat{\mu}_{(\mathrm{PGP})} = \frac{1}{m}\sum_{i=1}^{m}\frac{y_i}{n_i} = \tilde{\pi} \tag{4.17}$$

as in (4.7) and the PGP estimator then is defined as

$$\hat{\rho}^{(\mathrm{PGP})} = \frac{1}{\hat{\mu}_{(\mathrm{PGP})}\left(1-\hat{\mu}_{(\mathrm{PGP})}\right)}\left[\frac{1}{m}\sum_{i=1}^{m}\frac{y_i(y_i-1)}{n_i(n_i-1)} - \hat{\mu}^2_{(\mathrm{PGP})}\right] \quad . \tag{4.18}$$

A third case emerges when weights are proportional to the number of times a certain pair of observations occurs as a share of the total number of observations. Weights are thus defined as $w_i = \dfrac{1}{n(n_i-1)}$ yielding the estimator

$$\hat{\mu}_{(\mathrm{PPR})} = \frac{1}{n}\sum_{i=1}^{m} y_i = \hat{\pi} \quad . \tag{4.19}$$

With this configuration, the PPR estimator of ρ is given by

$$\hat{\rho}^{(\mathrm{PPR})} = \frac{1}{\hat{\mu}_{(\mathrm{PPR})}\left(1-\hat{\mu}_{(\mathrm{PPR})}\right)}\left[\frac{1}{n}\sum_{i=1}^{m}\frac{y_i(y_i-1)}{n_i-1} - \hat{\mu}^2_{(\mathrm{PPR})}\right] \quad . \tag{4.20}$$

As Ridout et al. (1999, 140) note, with equal cluster sizes there are classes of estimators which yield identical estimates. The first class is composed of $\hat{\rho}^{(\mathrm{AOV})}$, $\hat{\rho}^{(\mathrm{MAK})}$ and $\hat{\rho}^{(\mathrm{UBE})}$. The second class is made up of $\hat{\rho}^{(\mathrm{KEQ})}$ and $\hat{\rho}^{(\mathrm{KPR})}$ and the third class consists of $\hat{\rho}^{(\mathrm{FLC})}$, $\hat{\rho}^{(\mathrm{PEQ})}$, $\hat{\rho}^{(\mathrm{PGP})}$ and $\hat{\rho}^{(\mathrm{PPR})}$.

4.3.4 Estimators Based on Random Effects Models

Estimators based on the variance decomposition of a random effect model can of course also be specified for binary data. Rodríguez and Elo (2003) give an overview of the composition and estimation of intraclass correlation estimators based on random effect models and illustrate their use with the xt commands of STATA. They present estimators based on *probit*, *logit* and *complementary log-log* (*cll*) models. These types of models can also be estimated using package lme4 of R.

The basic underlying model to be estimated is of course identical to the one specified above and is repeated here just for convenience of reading:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij}.$$

With the left hand side of the above equation being a dichotomous variable, a link function ensures that the linear predictor fits the scale of the outcome. In probit

models, the probability for the outcome variable to take on the value of 1, given the random effect, is $\pi_{ij} = P(y_{ij} = 1|\alpha_i) = \Phi(\mu + \alpha_i)$ leading to the following model:

$$\Phi^{-1}(\pi_{ij}) = \mu + \alpha_i.$$

A nice feature of this model is its flexibility. It can be reformulated in terms of y^*_{ij} being a latent variable that serves as an indicator for the outcome which is positive iff the indicator is above a given threshold. Setting this threshold to zero and $\sigma^2_\epsilon = 1$ we can write the formula for the estimator of the intraclass correlation coefficient in the same way as in the case of a continuous variable:

$$\hat{\rho}^{(\text{RE}-\text{PRO})} = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + 1} \quad . \tag{4.21}$$

With logit models, observations given the random effect are assumed to have independent Bernoulli distributions with probabilities $\pi_{ij} = P(y_{ij} = 1|\alpha_i) = F(\mu + \alpha_i)$ where $F(\cdot)$ is the logistic distribution with c.d.f $F(\mu) = \frac{e^\mu}{1+e^\mu}$. Taking the inverse, F^{-1}, gives the logit model:

$$F^{-1}(\pi_{ij}) = \log \frac{\pi_{ij}}{1 - \pi_{ij}} = \mu + \alpha_i.$$

The construction of the estimator for the intraclass correlation coefficient follows the same basic logic as before. However, we now assume the errors, ϵ_{ij}, to follow a logisitc distribution with mean 0 and variance σ^2_ϵ. With the standard logisitc distribution, $\sigma^2_\epsilon = \pi^2/3 \approx 3.29$ we can construct the estimator for the intraclass correlation coefficient as

$$\hat{\rho}^{(\text{RE}-\text{LOG})} = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \frac{\pi^2}{3}} \quad . \tag{4.22}$$

A third type of model is the so called complementary log-log (cll) model where the link function is of the form $\log\left[\log(1 - \pi_{ij})\right] = \mu + \alpha_i$. Again, taking the inverse gives the log-Weibull distribution which is specified as

$$F(\mu + \alpha_i) = 1 - \exp\left[-\exp(\mu + \alpha_i)\right].$$

In the cll model the error terms are usually assumed to have "(reverse) extreme value distributions" (Rodríguez and Elo, 2003, 39). This distribution has cumulative density function $F(\epsilon_{ij}) = \exp\left[-\exp(e_{ij})\right]$ with mean ≈ 0.577 and variance $\pi^2/6$. Also for the cll model the setup of the estimator follows the same logic as before and is defined as

$$\hat{\rho}^{(\text{RE}-\text{CLL})} = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \frac{\pi^2}{6}} \quad . \tag{4.23}$$

The estimators of ρ for binary data investigated in Chapter 5 all are based on random effect models formulations of the logit type. Estimation of these types of models follows the same logic as before and basically the same techniques (ML and REML, specifically) can be used to estimate the variance components of the underlying random effects model. For a brief introduction into the foundations of generalized linear models (to which the above models belong) I have to refer to Chapter 7.3. This chapter also clarifies the pros and cons of each method and its specific effects on use of the variance components used for estimation of the intraclass correlation coefficient.

5 Monte Carlo Simulation Studies

In this chapter, the results of some Monte Carlo simulation studies on the behaviour of different estimation strategies for design effects and/or their components are presented. Both the design-based and the model-based approach are evaluated in terms of bias and precision of the respective estimators. This requires, as a first step, the generation of artificial universes in which a study variable is defined according to a specified model. The generation of these universe(s) is described in the first section (5.1) of this chapter. In Section 5.2 the general set-up of the simulation study is descried in more detail. Section 5.3 presents the results of the Monte Carlo estimation of the true design effect. In Section 5.4 design-based estimation of *deff* for the mean and the median (Section 5.4.2) are evaluated empirically. Section 5.5 presents the results of model-based estimation approaches and Section 5.6 synthesizes the advantages and disadvantages of either approach and compares them to each other. The last section (5.7) presents the results of a simulation which aims at a separation of design- and interviewer effects based on a nested random effects model.

5.1 Generation and Structure of Universes

A first step in every Monte Carlo simulation study is the generation of universes from which samples can be drawn. Elements of these universes are characterized by a) *structural* and b) *substantive variables*. Structural variables are pre-defined characteristics such as PSUs and interviewer cluster assignment indicators. Substantive variables (or *study variables*), on the other hand, are generated according to a predefined model (here: a random effects model) which includes the structural variables as predictors (e.g. PSUs as random effects). The distributional parameters of the study variable in the universes serve as a benchmark against which the simulation results are being evaluated.

One can distinguish between two classes of clustered universes which are generated according to a) a one-way ANOVA (see Section 7.3) and b) a nested two-way ANOVA (see Section 7.3) model. The generation of universes of the first class is described in Section 5.1.1, the generation of universes of the second class in Section 5.1.2.

5.1.1 Geographically Clustered Universes

Universes of the first class are created according to the common parameter model given in (2.9) on page 29. The generation of universes belonging to this class is motivated by a situation where clustering on the study variable is only due to geographical units (i.e. PSUs). The number of geographical clusters in the universe is chosen $M = 1\,000$ with $N_i = 500$, $i = 1, \ldots, M$, elements in each geographical cluster. Multiple universes are generated to account for different levels of homogeneity. The levels of ρ are chosen $\{0.01, 0.02, 0.03, 0.04, 0.05, 0.10, 0.15, 0.20\}$. When necessary, universes are labelled U_ρ to distinguish between them. The vector of the study vari-

able, y, is generated according to the common parameter model given in Valliant et al. (2000). This model assumes the response to be given by

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad , \tag{5.1}$$

as in (9) on page 156. In the generation of the universes, the FR estimator of ρ implemented in the deff() function of the Hmisc package was used to evaluate ρ in the respective population. A universe is accepted if the estimated value of ρ is in the interval $\rho \pm \rho/1000$. The overall mean of the study variable was chosen zero.

In addition, a further set of universes is generated for *binary outcome data*. In that case, an additional parameter, namely the the overall rate of success, π, must be considered as a factor of the simulation study. Data are generated for three different levels: $\pi = \{0.05, 0.25, 0.50\}$. The universes are then created assuming a common-correlation model with an underlying beta-binomial distribution: This was done following the procedure described in Ridout et al. (1999, 138). Here, the AOV estimator of ρ is used to evaluate ρ in the population. Again, only if ρ lays within the boundaries of the interval defined by $\rho \pm \rho/1000$ the universe was accepted.

5.1.2 Universes with a Nested Structure

In addition to geographically clustered universes, a second class of populations is generated in which additional homogeneity is introduced by interviewers nested within PSUs. Thus, we must distinguish between ρ_{PSU} and ρ_{INT}, indicating homogeneity due to geographical clustering (namely through PSUs) and homogeneity due to the interviewer. A further parameter to consider within this setting is the share of the variation that is attributed to interviewer clustering, denoted by s_{INT}.

Due to reasons of computation time, only four different levels of the magnitude of ρ_{PSU} and ρ_{INT} are investigated in the simulation studies, respectively. These are $\rho_{\text{PSU}}, \rho_{\text{INT}} = \{0.02, 0.05, 0.10, 0.20\}$. Three different levels of s_{INT} $(0.33, 0.50, 0.67)$ are assumed, indicating that 33%, 50% and 67% of the total variance explained is attributed to the interviewer level. These parameter combinations result in $4 \times 4 \times 3 = 48$ additional universes. Figure 8 illustrates the nested structure of the universes.

Interviewers are assumed to be nested within PSUs, with exactly $I = 2$ interviewers operating in each cluster (Gabler and Lahiri, 2009). Thus the total number of interviewer clusters in the population is $K = M \cdot I = 1\,000 \cdot 2 = 2\,000$. The interviewer clusters are generated perfectly balanced, this means each is of the same size in the population, $N_{ik} = 250$, with $k = 1, \ldots, I$. With the jth element of the kth interviewer cluster in the ith PSU is associated a value on the study variable, y_{ikj}, $j = 1, \ldots, N_{ik}$. Figure 8 illustrates the basic structure of the nested universe.

The response variable, y, is generated stepwise: First, homogeneity due to geographical clustering is generated in y_{PSU} according to the common mean model given in Valliant et al. (2000), at a level of tolerance of $\pm\rho_{\text{PSU}}/100$ due to computation time evaluated based on the FR estimator of the deff() function. In a second step, interviewers nested in PSUs are assumed as elements introducing clustering and the common mean

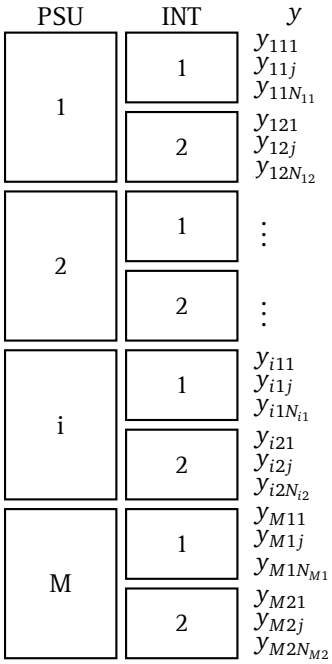


Figure 8: *Illustration of nested universe structure*

model is again applied to generate a second outcome variable, y_{INT}, in a procedure identical to the one of the first step. Estimated population ρ must fulfil the tolerances as before. Then, y_{PSU} and y_{INT} are combined so that they yield the final response variable

$$y = y_{\text{PSU}} \cdot w_{\text{PSU}} + y_{\text{INT}} \cdot w_{\text{INT}} \quad , \tag{5.2}$$

where $w_{\text{INT}} = 1 - w_{\text{PSU}}$. w_{PSU} and w_{INT} are chosen such that $\frac{\sigma_\alpha}{\sigma_\alpha + \sigma_\beta + \sigma_\epsilon} = s_{\text{INT}}$ where σ_α and σ_β are the variances of the random effects of the following model (see also Section 7.3):

$$y_{ikj} = \mu + \alpha_i + \beta_{ik} + \epsilon_{ikj} \quad . \tag{5.3}$$

The parameters of the model are estimated by the lmer() function of R's lme4 package and variance components are being extracted using the VarCorr() function. This process is repeated until also s_{INT} lies within the interval $s_{\text{INT}} \pm s_{\text{INT}}/100$.

5.2 Aim and Design of the Monte Carlo Simulation Studies

In the following, the word *scenario* shall be used to refer to a specific combination of certain parameters of the universe, for example drawing type (equal vs. unequal cluster sizes and hence equal vs. unequal inclusion probabilities of elements of different PSUs), population ρ and m (number of clusters drawn or, equivalently, $\bar{b}$, the average cluster size). For example, the combination of $\rho = 0.02$ and $m = 150$ under

equal probability cluster sampling is referred to as a scenario. A *setting*, on the other hand, is a combination of scenarios, e.g. the set of 12 (4 levels of $\rho \times 3$ levels of m) scenarios under cluster sampling with equal probabilities.

In what follows, certain properties of estimators (e.g. their average, standard deviation, relative bias, etc.) will be analysed given a specific scenario. Furthermore, comparisons between scenarios are made and systematic effects of single parameters of the setting or of combinations of parameters are extracted. Finally, when possible, comparisons of common effects within settings are compared to those of other settings, e.g. the effects of allowing for variation in cluster sizes and hence switching from equal to unequal probability cluster sampling which, in essence, requires design weighting at the estimation stage.

5.2.1 Simulation Strategy

The number of clusters drawn, m, under cluster sampling is either 150, 300 or 500. As a further source of variation, the cluster size may either be equal or unequal – implying of course also equal or unequal inclusion probabilities with fixed cluster sizes in the population ($N_i = N_\bullet = 500$). We shall refer to the scenario with fixed and equal cluster sizes in the sample to *two-stage cluster sampling with equal cluster sizes* (clu2e) and to the other scenario as *two-stage cluster sampling with unequal cluster sizes* (clu2u)[11]. In either scenario selection of PSUs is done by srs, thus the selection probabilities of all PSUs, π_i, are equal. At the second stage, elements within selected PSUs are also sampled using srs. The sample size, n, under clu2e is guaranteed to be 3 000 as

$$n_i = \begin{cases} 20 & \text{if } m = 150 \\ 10 & \text{if } m = 300 \\ 6 & \text{if } m = 500. \end{cases}$$

Thus, under clu2e, also the inclusion probabilities on the second stage are constant: $\pi_{j|i} = \dfrac{n_i}{N_i} = \dfrac{c}{500}$, $j = 1, \ldots, 500$, i.e.

$$\pi_{j|i} = \begin{cases} \frac{20}{500} = 0.04 & \text{if } m = 150 \\ \frac{10}{500} = 0.02 & \text{if } m = 300 \\ \frac{6}{500} = 0.012 & \text{if } m = 500. \end{cases}$$

When allowing for variation in cluster sizes in the sample, this variation is chosen such that the coefficient of variation of cluster sizes is constant. This ensures comparability since variation of estimators is not influenced by different levels of variation in weights. Cluster sizes are sampled from a uniform distribution. Thus, interval boundaries, $[a, b]$, must be chosen such that $cv(n_\bullet) = c = \frac{\sqrt{\frac{1}{12}(b-a)}}{\frac{1}{2}(a+b)}$ for every

11 For more details on sample designs see also Section 2.1

$m = \{150, 300, 500\}$. This criterion is satisfied if we sample n_i with $E(n_i) = \frac{n}{m}$ from a symmetric distribution (here: a uniform distribution) with boundaries

$$[a, b] = \begin{cases} [10, 30], & \text{if } m = 150 \\ [5, 15], & \text{if } m = 300 \\ [3, 9], & \text{if } m = 500 \end{cases} . \tag{5.4}$$

The fact that the number of sampled elements per PSU, n_i, varies from cluster to cluster under clu2u implies that second stage inclusion probabilities, $\pi_{j|i}$, will vary, increasing the variance of the HT estimator of the population mean (see equation (2.5)). The $\pi_{j|i}$ under clu2u will lay in the interval

$$\pi_{j|i} = \begin{cases} \left[\frac{10}{500} = 0.02, \frac{30}{500} = 0.06\right] & \text{if } m = 150 \\ \left[\frac{5}{500} = 0.01, \frac{15}{500} = 0.03\right] & \text{if } m = 300 \\ \left[\frac{3}{500} = 0.006, \frac{9}{500} = 0.018\right] & \text{if } m = 500. \end{cases}$$

It is obvious that the expected value of $\sum_{i=1}^{m} n_i$ is $E(n_i)m = 3\,000$ for all m.

In the nested universes every element in the population is a priori associated with an interviewer. That is, if the jth element of the ith selected PSU is selected into the sample of size n_i, the size of the respective interviewer clusters, n_{ik}, depends on which elements j have been selected. Since selection of elements within a PSU is done by srs, the expected size of the kth interviewer cluster in the ith PSU is $\bar{b}_{ik} = {}^{n_i}/_2$ due to the fact that in the population there are two evenly large ($N_k = 250$) interviewer clusters assigned to every PSU.

For the simulation studies, a set of 10 000 *sample vectors* is generated for each level of m and for each drawing type (clu2e and clu2u). The sample vectors are logical and indicate whether or not a population element is a member of the sample.

Labelling of scenarios and settings is as follows: any quantity will be referred to by the following index scheme: $Var\left(\hat{\theta}_{(\text{sd})[\text{cs}]\langle m\rangle\{\rho\}}\right)$ with *sd* denoting *sample design* and *cs cluster size type*. Following this scheme, the variance of $\hat{\theta}$ for a two-stage cluster sample with equal cluster sizes, 300 PSUs drawn from a clustered population and with population parameter $\rho = 0.05$ would be denoted by $Var\left(\hat{\theta}_{(\text{clu2})[\text{eq}]\langle 300\rangle\{0.05\}}\right)$. Whenever multiple values of a parameters are under consideration, the plus sign (+) is used as an indicator. In many figures, for example, we will restrict the illustration to only four selected levels of ρ, namely 0.02, 0.05, 0.10, and 0.20. After mentioning the subset under consideration in the text, this setting, for example, will be referred to as $Var\left(\hat{\theta}_{(\text{clu2})[\text{eq}]\langle 300\rangle\{+\}}\right)$ for simplicity.

5.3 Monte Carlo Estimation of the True Design Effect

When taking the design-based perspective on estimation of design effects, the simulation study first permits us to estimate the empirical variances of the HT estimator

under cluster sampling and simple random sampling. Then, the empirical variance under cluster sampling has to be divided by the empirical variance under srs, this ratio being the design effect[12] by the definition of Kish (1965).

For the sake of clarity, I will illustrate the set-up of the simulation studies that underlay this section in detail: The simulation study based on continuous data has 2 (type of inclusion probabilities) × 3 (no. of clusters) × 8 (levels of ρ)=48 factors. The combinations of factors, or scenarios, are summarized in Table 19 in the appendix. In addition, the Monte Carlo estimated true design effect is also estimated for binary data. For binary outcome data, the combinations of 144 further scenarios are summarized in Table 19 (also in the appendix).

In each of the scenarios, 10 000 samples are drawn from the respective population and the HT estimator is being estimated both based on a clustered sample and on a simple random sample of size 3 000. Thus, this setting yields a total of $(48 + 144) \times 10\,000 = 1\,920\,000$ simulation results.

5.3.1 Continuous Data

This subsection describes the Monte Carlo estimation of the true design effect with continuous data. The study variable in the universe was generated as described in Section 5.1. As we concentrate on precision of the HT estimator a slight skewness in the study variable in some of the universes was accepted. In those cases, this may lead to plots that are symmetric but not around grand mean of other universes. However, this does not disturb the conclusions to be drawn from the simulations.

5.3.1.1 Two-Stage Cluster Sampling with equal Cluster Sizes

In this subsection, we will consider the case of two-stage cluster sampling with an equal number of samples elements per cluster (clu2e) with continuous data. The fact that an equal number of elements, $n_i = \frac{3000}{m}$, per PSU is drawn randomly from clusters which are of equal size, $N_i = 500$, in the population, ensures equal inclusion probabilities and makes weighting unnecessary.

Under this setting, $m = \{150, 300, 500\}$ clusters are drawn by srs from the population of $M = 1\,000$ clusters. Then, on the second stage, a sub-sample of either 20, 15 or 6 ultimate sample elements is chosen by srs from each cluster. Each sample that can be drawn under this design is of fixed size $n = 3\,000$.

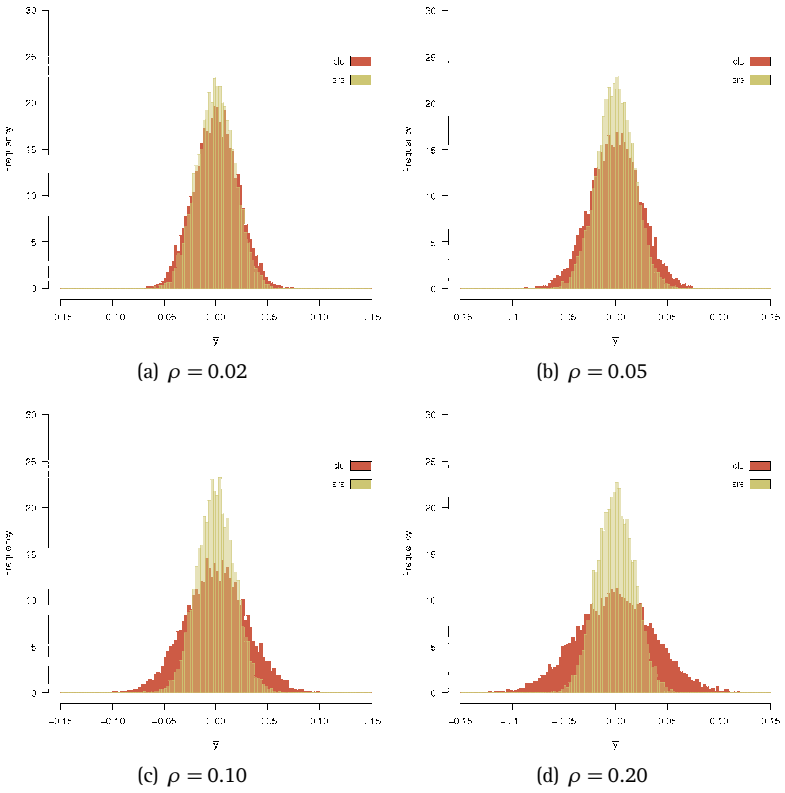
Additionally, a srs of ultimate sample units of the same size is drawn directly from the population. Based on either sample, the HT estimator of the population mean (see formula (2.2) on page 27) is calculated[13]. This procedure is repeated 10 000 times for each combination of ρ and m.

An illustration of the distribution of the 10 000 values of the HT estimator is given in Figure 9 which graphically displays the variance of the HT estimator under two-

12 see (3.2) on page 36

13 Due to the fact that with equal cluster sizes inclusion probabilities do not vary, the estimator simplifies to the usual sample mean, given by $\hat{\bar{y}} = n^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n_i} y_{ij}$.

stage cluster sampling and srs for $m = 150$ (i.e. $\bar{b} = 20$) for selected values of ρ, hence $Var\left(\hat{\bar{y}}^{(HT)}_{(\text{clu2})\,[\text{eq}]\,\langle 150\rangle\,\{+\}}\right)$.



(a) $\rho = 0.02$ (b) $\rho = 0.05$

(c) $\rho = 0.10$ (d) $\rho = 0.20$

Figure 9: *Overlaid Histograms based on a simulation of* 10 000 *repeated draws under srswor and two-stage equal probability cluster sampling (*$m = 150$*)*

What can be seen at first glance is that the histogram in lighter colour in front – displaying the distribution of 10 000 HT estimates based on simple random sampling of ultimate sample elements – has a) identical shape in each plot and b) less fat tails than the darker histogram in the background. As each histogram visually depicts the empirical variance of the HT estimator for a specific sample design, the overdispersion of the histogram in the background indicates that the design-based design effect, defined as the ratio $\dfrac{Var\left(\hat{\bar{y}}_{(\text{clu2e})}\right)}{Var\left(\hat{\bar{y}}_{(\text{srs})}\right)}$, in all scenarios exceeds unity. What can also be seen is that with increasing homogeneity in the population the tails of the histogram in the background become fatter which indicates higher variation in the empirical distribution of estimates of $\hat{y}^{(HT,clu2,equal)}$.

Another way to look at the variation of the HT estimator in given scenarios is through the grouped boxplots displayed in Figure 10. Here, the lower boxplot of

each panel – indicating the distribution of 10 000 HT estimates under clu2 with equal cluster sizes – is wider than the accompanying boxplot at the top of the panel. This indicates the increased variance of the HT estimator under two-stage cluster sampling with equal cluster sizes as compared to its variance under simple random sampling.

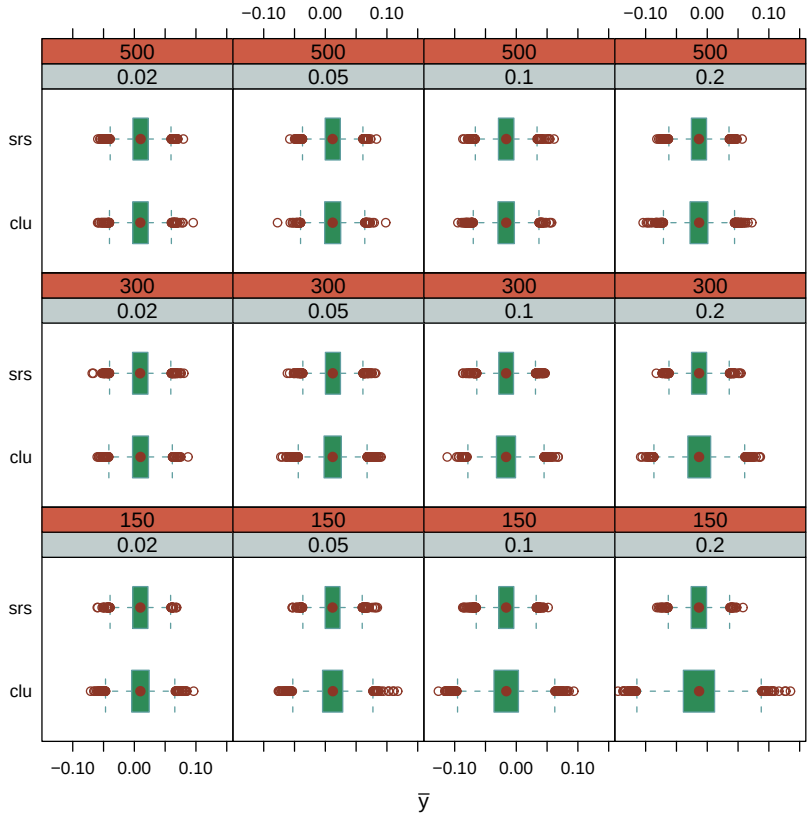


Figure 10: *Grouped boxplots of the distribution of the HT estimator under srs and two-stage cluster sampling with equal cluster sizes for given scenarios with continuous data*

What can easily be seen from Figure 10 is that both, ρ and m (and hence $\bar{b}$) have an influence on the magnitude of the estimated design effect. With large cluster sizes the effect (in absolute terms) of an increase in ρ is of course greater than with small cluster sizes[14]. This is even more obvious in the following Figure (11). For sake of completeness, the variance of the HT estimator under clu2 and srs as well as the Monte Carlo estimated true design effect based on these quantities for each scenario is given in Table 21 in the appendix.

14 Note that the non-linear shape of the dots is due to a non-linear increase of the two lower most levels of the model parameter ρ on the y-axis.

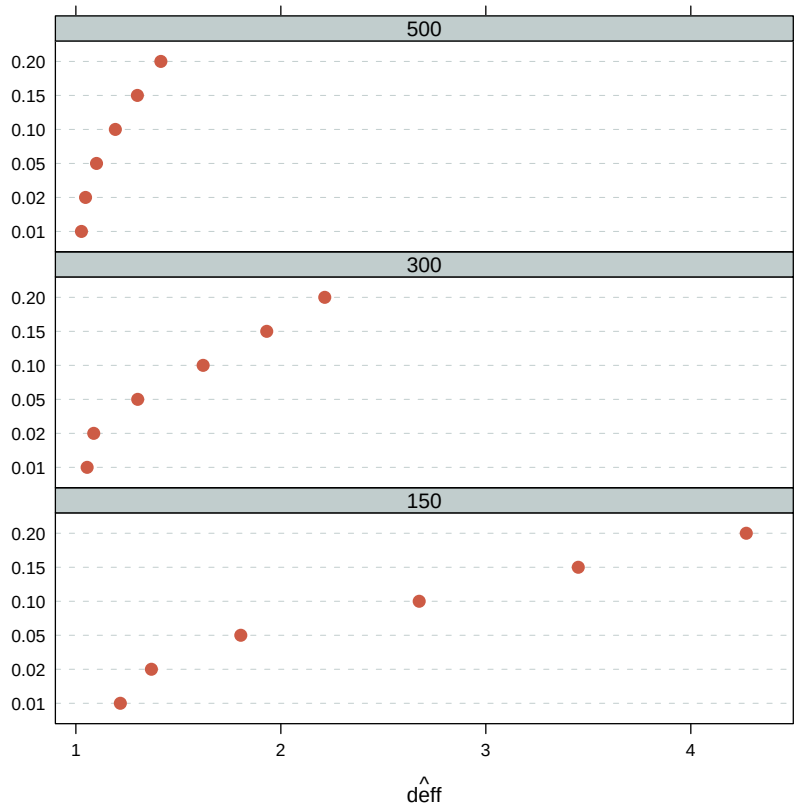


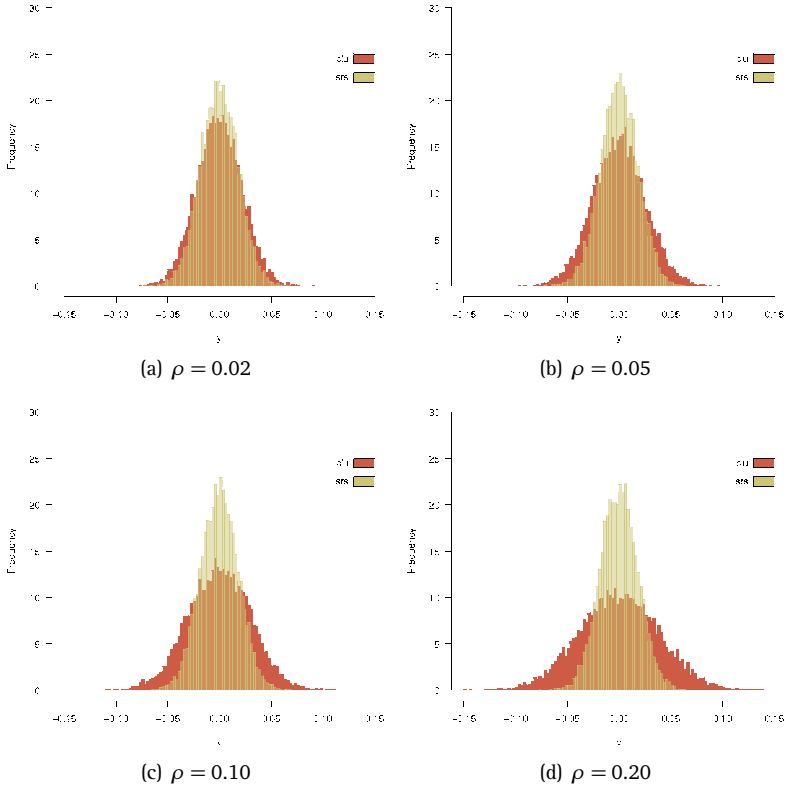
Figure 11: *Grouped dotplots of the Monte Carlo estimated true design effect for two-stage cluster sampling with equal cluster sizes for given scenarios with continuous data*

5.3.1.2 Two-Stage Cluster Sampling with unequal Cluster Sizes

This scenario is similar to the one described in the previous section with the difference that now cluster sizes in the sample are allowed to vary (clu2u) within the boundaries defined in (5.4). This variation in the number of selected individuals per PSU implies that design weights, $w_i = {}^{N_i}/_{n_i}$, will vary. This additional variation is reflected in the variance of the HT estimator (see (2.5) and (2.6) in Section 2.2). Besides, all further steps of the simulation set-up and the analytical procedure are the same as described in the previous subsection. The following overlaid histograms show the variance of the estimator under clu2u and srs for $m = 150; \bar{b} = 20$ and selected values of ρ.

The shape and general interrelation of the histograms are similar to the scenario with equal cluster sizes. However, the distribution of HT estimates under cluster sampling is even wider than in the equal cluster sizes scenario, reflecting the aforementioned increase of variance in the estimator due to additional variance introduced by variation in weights.

Taking a look on the distribution of HT estimates based on a wider range of scenar-



(a) $\rho = 0.02$ (b) $\rho = 0.05$

(c) $\rho = 0.10$ (d) $\rho = 0.20$

Figure 12: *Overlaid Histograms based on a simulation of* 10 000 *repeated draws under srs and two-stage cluster sampling with unequal cluster sizes (*$m = 150$, $\bar{b} = 20$*)*

ios (figure 13) we can, again, see that the influence of the average cluster size rules this distribution – however, now generally on a higher level as additional variation of the HT estimator is introduced by the design weights[15]. Also with cluster sampling with unequal cluster sizes (and hence unequal inclusion probabilities), the influence of ρ heavily depends on the average cluster size as can be seen from Figure 14. With large average cluster sizes an increase in ρ has a larger effect on the magnitude of $\widehat{deff}$ than in a scenario with smaller average cluster sizes. The plots indicate an increase in variance of the HT estimator due to additional variation introduced by design weights. This becomes even more obvious in Table 22 in the appendix which is composed similar to the one before but now displays the variance of the HT estimator under two-stage sampling with unequal cluster sizes and under srs along with the Monte Carlo estimated true design effect defined as the ratio of these two quantities.

15 This, of course, has an effect only on the spread of the distribution of the HT estimator under cluster sampling and not under srs.

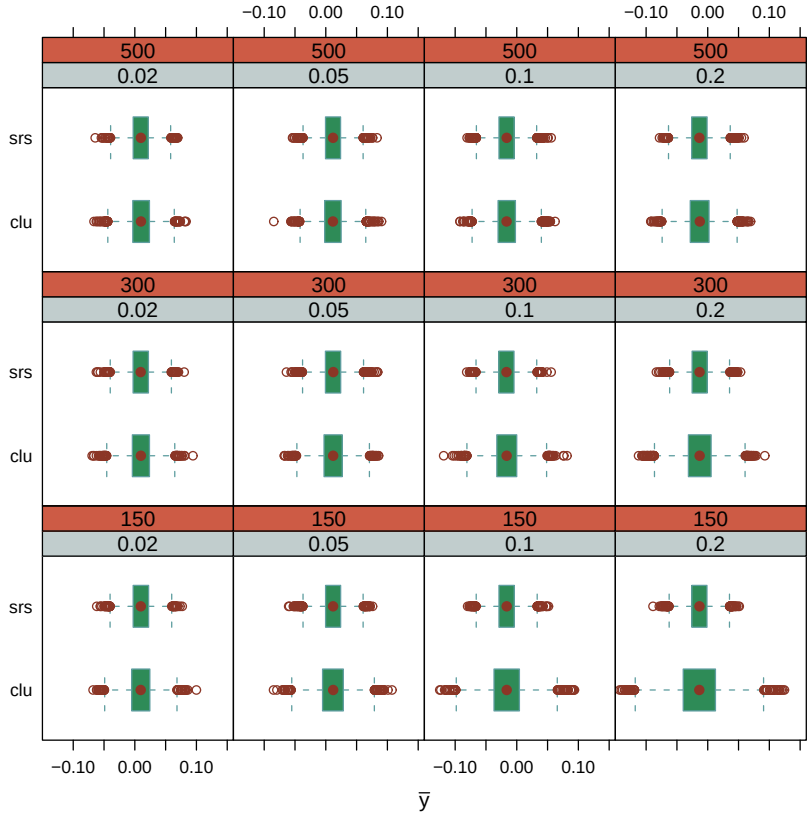


Figure 13: *Grouped boxplots of the distribution of the HT estimator under srs and two-stage cluster sampling with unequal cluster sizes for given scenarios with continuous data*

5.3.1.3 Comparison of Two-Stage Cluster Sampling with equal and unequal Cluster Sizes

Unequal cluster sizes imply unequal inclusion probabilities and these, in turn, require use of design weights for unbiased estimates. The HT estimator considers design weights and as such serves as an unbiased estimator for the population mean. Including weights, however, increases the variance of the estimator. This inflation of variance can already be seen by comparing Tables 21 (equal cluster sizes) and 22 (unequal cluster sizes) where each entry in the fourth column of Table 22 is larger than the corresponding entry of Table 21. This is illustrated graphically in Figure 15.

Here, every dot in the upper part of a given panel (i.e. indicating unequal cluster sizes and hence unequal inclusion probabilities) lays more to the right than the corresponding dot of the lower part of the panel. This indicates the aforementioned fact that the variance of the HT estimator as well as the design effect is larger for cluster sampling with unequal cluster sizes.

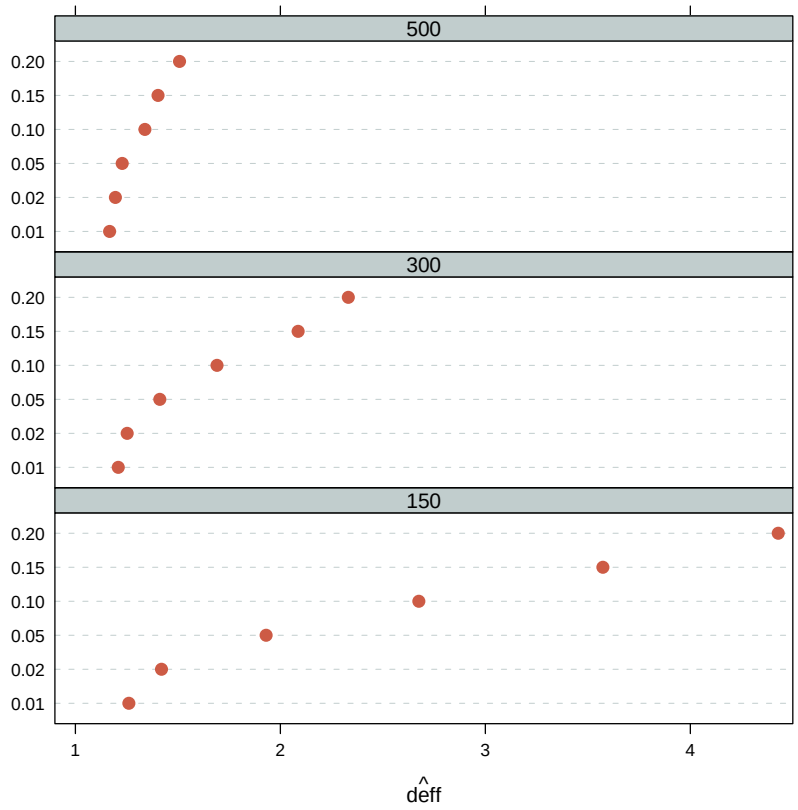


Figure 14: *Grouped dotplots of the Monte Carlo estimated true design effect for srs and two-stage cluster sampling with unequal cluster sizes for given scenarios with continuous data*

The magnitude of the (model-dependent) ratio $\frac{Var\left(\hat{\bar{y}}^{(HT)}_{(\text{clu2})\,[\text{ue}]\,\langle+\rangle\,\{+\}}\right)}{Var\left(\hat{\bar{y}}^{(HT)}_{(\text{clu2})\,[\text{eq}]\,\langle+\rangle\,\{+\}}\right)}$ is given in Table 4.

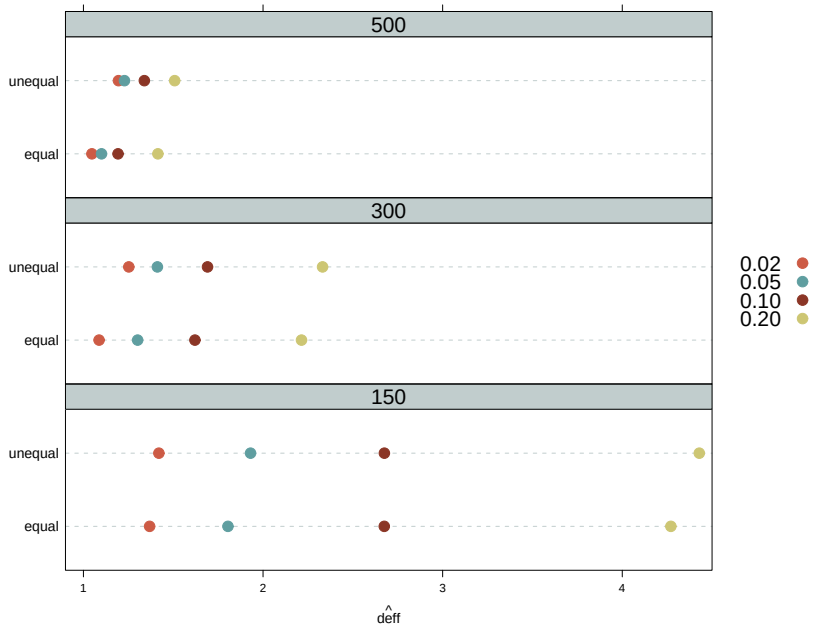


Figure 15: *Dotplot of mean $\widehat{deff}$ for levels of ρ and average cluster sizes*

Table 4: *Summary of the simulation study with two-stage cluster sampling with equal and unequal selection probabilities*

ρ	m	$\frac{Var\left(\hat{\bar{y}}^{(HT)}_{(\text{clu2})\,[\text{ue}]\,\langle+\rangle\,\{+\}}\right)}{Var\left(\hat{\bar{y}}^{(HT)}_{(\text{clu2})\,[\text{eq}]\,\langle+\rangle\,\{+\}}\right)}$
0.01	150	1.0714
0.01	300	1.1307
0.01	500	1.1449
0.02	150	1.0629
0.02	300	1.1403
0.02	500	1.1162
0.03	150	1.0808
0.03	300	1.0797
0.03	500	1.1250
0.04	150	1.0704
0.04	300	1.1335
0.04	500	1.1324
0.05	150	1.0686
0.05	300	1.0828
0.05	500	1.1035
0.10	150	1.0115
0.10	300	1.0575
0.10	500	1.1069
0.15	150	1.0442
0.15	300	1.0632
0.15	500	1.0872
0.20	150	1.0456
0.20	300	1.0483
0.20	500	1.0819

5.3.2 Binary Data

The distributions with binary data are very similar to the ones of the previous section. With dichotomous variables, however, an additional parameter comes into play, namely the overall rate of success of the study variable, π, say. Hence, samples are drawn from three different populations with $\pi = \{0.05, 0.25, 0.50\}$ described earlier (see section 5.1).

5.3.2.1 Two Stage Cluster Sampling with equal Cluster Sizes

Estimation of the true design effect as defined earlier is very similar to the continuous data case. All we need is an appropriate estimator for the overall rate of success, π. This is where ratio estimation comes into play. Inspecting Figure 16, we can see

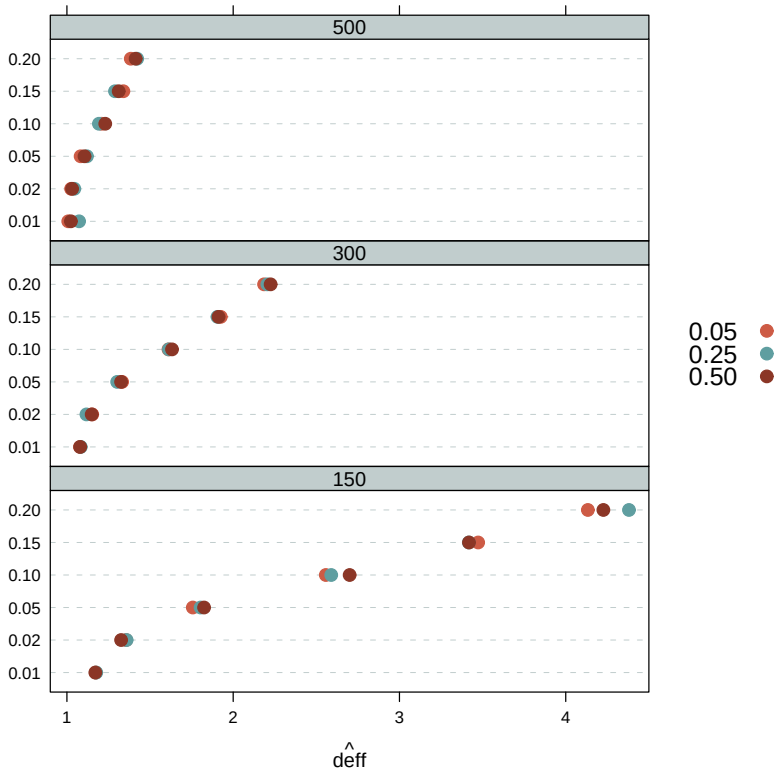


Figure 16: *Grouped dotplots of mean $\widehat{deff}$ under cluster sampling with equal cluster sizes for given scenarios with binary data*

that in most scenarios, the Monte Carlo estimated true design effect increases with π – nevertheless, generally, there is only little variance. Again, as we saw earlier, its magnitude is more influenced by cluster size than by population ρ.

5.3.2.2 Two Stage Cluster Sampling with unequal Cluster Sizes

When cluster sizes vary and design weighting comes into play, also with binary data Monte Carlo estimation of the true design effect is affected by the additional variance introduced by the weights. However, the basic picture stays the same as before as the magnitude of the design effect is depending mainly on the average cluster size and not so much on population ρ. This can be seen from Figure 17 which is similar in structure to Figure 16 of the previous section. As with cluster sampling with equal

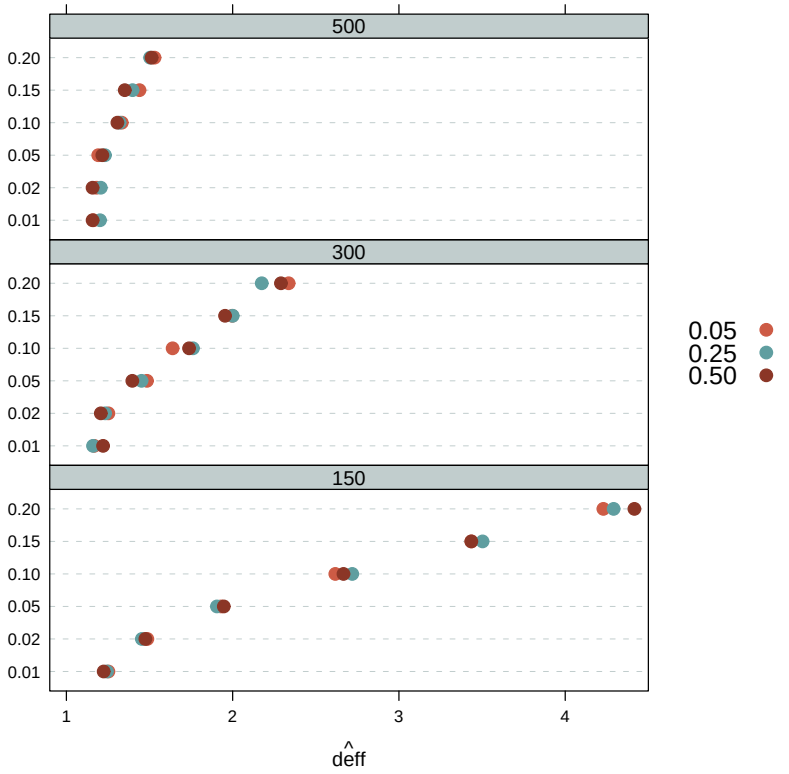


Figure 17: *Grouped dotplots of mean $\widehat{deff}$ under cluster sampling with unequal cluster sizes for given scenarios with binary data*

cluster sizes, the Monte Carlo estimated true design effect tends to be largest for π for any given scenario – however this tendency is not so obvious as in the previous setting.

5.3.2.3 Comparison of Two-Stage Cluster Sampling with equal and unequal Cluster Sizes

When comparing cluster sampling with equal and unequal cluster sizes (and hence unweighted and weighted estimation), we can observe similar structure with dichotomous outcome as with continuous data. Also with binary data, the Monte Carlo estimated true design effect is larger for cluster sampling with unequal than for equal cluster sizes as design weighting introduces additional variance to the estimates. This is illustrated in Figure 18 where we can see that each dot, representing the estimated design effect under a given scenario, in the upper part of a panel is further to the right than the corresponding dot in the lower part. In this direct comparison, we can also

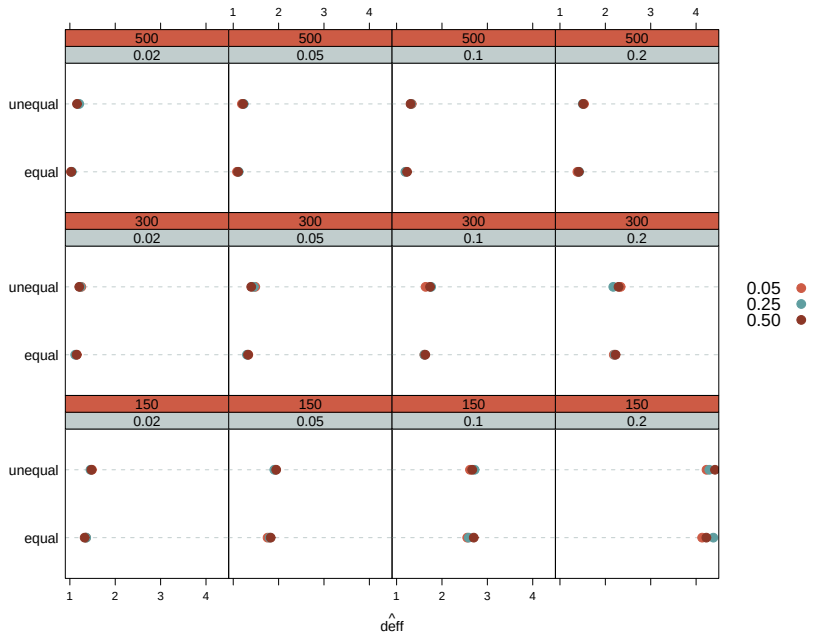


Figure 18: *Grouped dotplots of $\widehat{deff}$ under cluster sampling with equal vs. unequal cluster sizes for given scenarios with binary data*

observe the tendency of the estimators to be more sensitive to variations in π under weighting as the dots in the upper part of a panel tend to be more widely spread than the ones in the lower half of the panel.

5.3.3 Comparing Estimation of the Monte Carlo estimated True Design Effect with Continuous and with Binary Data

As we have observed a tendency of the Monte Carlo estimated true design effect, in the binary setting, to vary for different levels of π, it is interesting to investigate whether this variation has any systematic effect as compared to the setting with continuous data. Figure 19 summarizes the relative deviation of the Monte Carlo estimated true

design effect for binary data, $\widehat{deff}_{[bin]}$, from Monte Carlo estimated true design effect for continuous data, $\widehat{deff}_{[con]}$, in units of $\widehat{deff}_{[con]}$. What can be seen at first glance

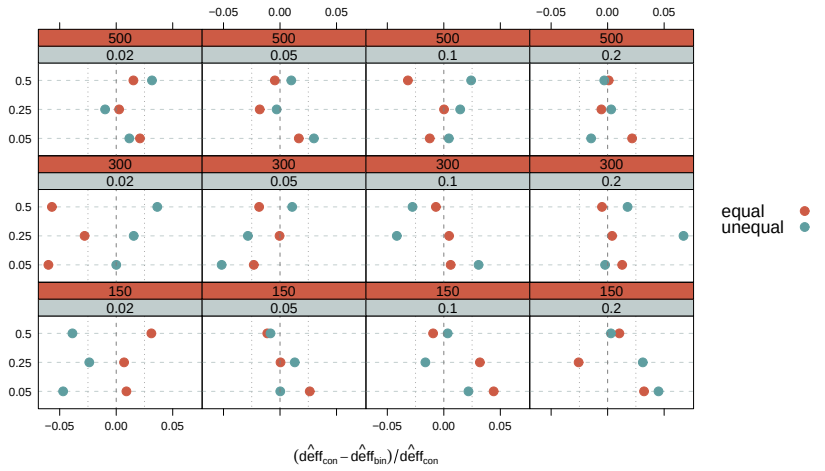


Figure 19: *Grouped dotplots of the relative deviance of the Monte Carlo estimated true design effect for binary data under cluster sampling with equal vs. unequal cluster sizes for given scenarios*

is that the deviations are relatively small, hardly any exceeding ±5%, most of them ranging within ±2.5%. There is a tendency of relative deviation to increase (in absolute value, i.e. it may change sign from - to +) with overall π. However, this effect is more obvious for small to medium than for large population ρ. All in all, deviations are smaller for cluster with equal than with unequal cluster sizes.

5.4 Design-based Estimation of the Design Effect

Strategies for design-based estimation of the design effect build upon variance estimation methods suitable for the estimator and the design under consideration (see Section 3.3.1). The design effect is interpreted as the ratio of the variance of an appropriate estimator under a given sample design to the variance of an appropriate estimator under srs (Kish, 1965). In real-world sampling practice, both estimators are being calculated on basis of a given complex sample. A huge variety of variance estimation methods has been proposed in the literature (see Section 3.3.1). The estimators of *deff* based on Taylor linearisation (Section 3.3.1.1) and JRR (Section 3.3.1.2) methods are evaluated in the following simulation studies. For the estimation of the design effect under two-stage cluster sampling, 10 000 samples are drawn from the different populations. Then, with every given sample the design effect is estimated using the variance of the unweighted and weighted sample mean of each of the aforementioned methods as enumerator of equation (3.2) according to (3.14) and (3.16). The variance of the sample mean under srs in the denominator of te respective equations is estimated treating the cluster sample data as if it was drawn using srs.

5.4.1 Continuous Data

This simulation setting builds upon the same universes that were used before. Also, the same sample vectors are used as in the previous setting. In fact, all simulation scenarios base upon the same set of 10 000 samples drawn from the set of different universes (see section 5.1 for more details). This enables direct comparisons between different scenarios and settings afterwards.

5.4.1.1 Cluster Sampling with equal Cluster Sizes

As two-stage cluster sampling with equal cluster sizes is a totally balanced design, the design-based estimators can be assumed to be highly precise. Precision, however, must be measured in a design-based manner since there is no model-based predicted value to evaluate the estimator against. Hence, in the following, the *coefficient of variation* (*cv* for short) shall serve as a measure of precision. The *cv* was chosen to make comparisons between different scenarios meaningful since it is defined as the standard deviation independent of the expected value of the estimator. The following dotplot shows the behaviour of the JRR and Taylor series estimator for given scenarios in terms of relative variation as measured by $cv(\widehat{deff})$. As can be seen at first glance,

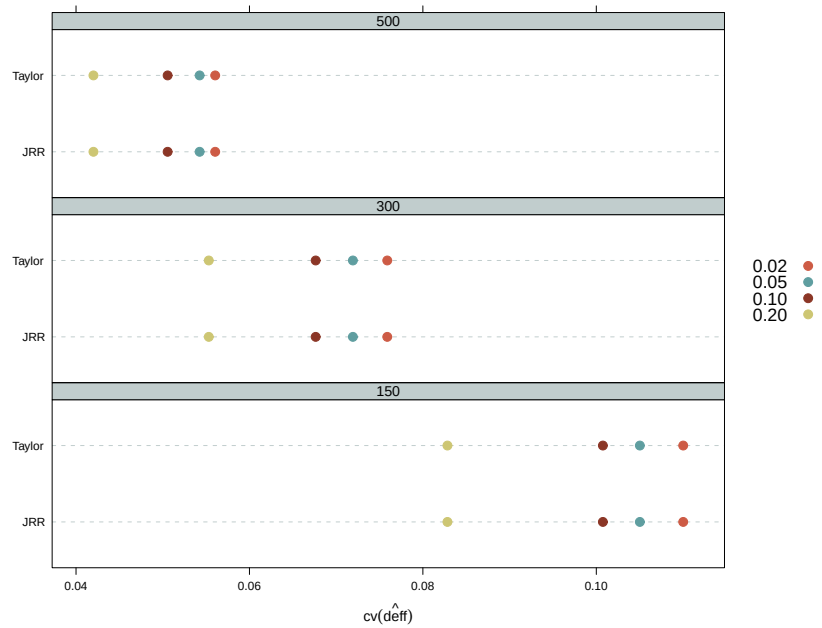


Figure 20: *Grouped dotplots of the cv of* $\widehat{deff}$ *under cluster sampling with equal cluster sizes for given scenarios with continuous data*

there is hardly any variation in terms of precision between estimators at a given level of population ρ and a given cluster size. Both, JRR and Taylor are equally precise in terms of their *cv*. There is, however, variation between levels of ρ. Both estimators

are less precise in scenarios where intraclass correlation is low. As the design effect in magnitude mainly depends on the average cluster size, the effect of variations between levels of m in cv is larger than the effect of variations in ρ.

This picture, however, can be misleading as it ignores the absolute magnitude of $\widehat{deff}$ yielded by the estimators. If we take a look at Figure 21, the scatter plot indicates that the JRR estimator produces values larger than the ones produced by the Taylor series estimator. In fact, with equal cluster sizes the JRR estimator yields values larger

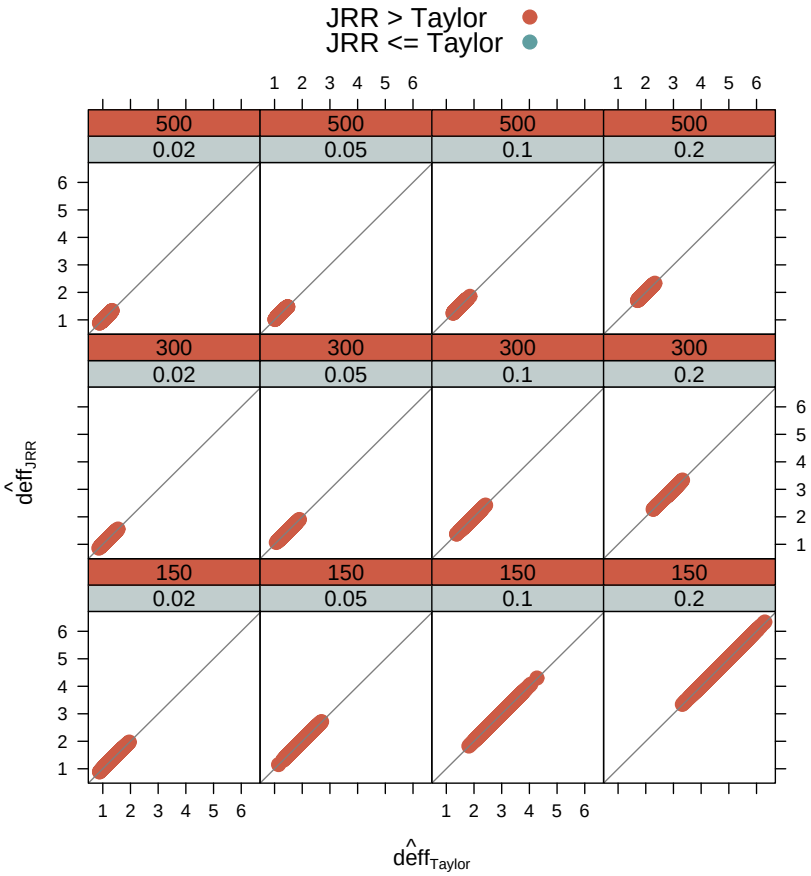


Figure 21: *Grouped scatter plots of JRR vs. Taylor series estimates of $\widehat{deff}$ under cluster sampling with equal cluster sizes for given scenarios with continuous data*

than those produced by the Taylor series estimator for a given sample in all of the 48 000 distinct samples of this setting. The question, then, is which of the estimators is less biased and estimates the Monte Carlo estimated and the model-expected true design effect most closely. Figure 22 gives an overview of the distribution of the estimates produced by JRR and Taylor estimators in different scenarios. What can be seen is that both estimators are downwards biased for the Monte Carlo estimated true design effect (red vertical line) and upwards biased for the model-expected design

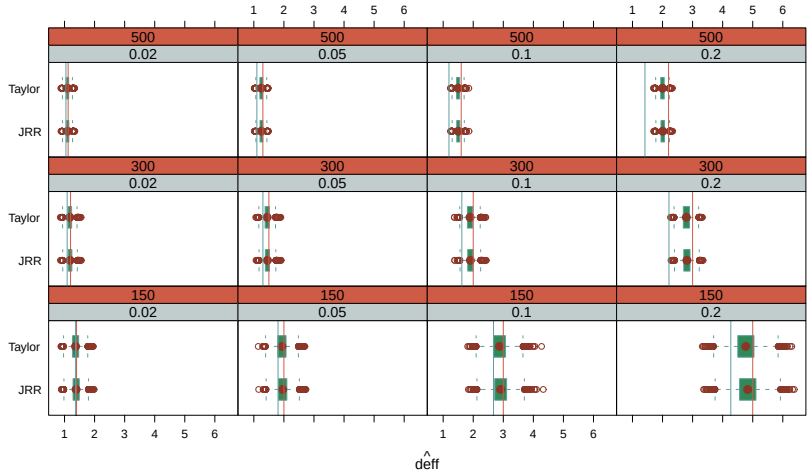


Figure 22: *Grouped boxplots of JRR and Taylor series estimates of $\widehat{deff}$ under cluster sampling with equal cluster sizes for given scenarios with continuous data*

effect (blue vertical line). This bias is influenced both by the level of the population parameter ρ and by the cluster size: High values of ρ and small cluster sizes generally increase bias.

5.4.1.2 Cluster Sampling with unequal Cluster Sizes

With unequal cluster sizes and hence unequal inclusion probabilities, design weighting must be incorporated into the HT estimator and hence also into its variance estimator. This, in turn, increases the variance of the estimator. The variance of the estimator for the denominator stays unchanged as it treats the data as if it had arisen from a srs design. It is obvious that this also increases the variance of the design-based estimators of the design effect but leaves the basic relationship between the estimators unchanged as can be seen in Figure 23.
Turning to deviation from the Monte Carlo estimated true design effect, we can observe the same change in the magnitude of the estimates JRR and Taylor series estimators yield as Figure 24 indicates.
However, there is no change in how the estimators relate to each other. As before, they are both positively biased for the Monte Carlo estimated true design effect.

5.4.1.3 Comparison of Two-Stage Cluster Sampling with equal and unequal Cluster Sizes

When switching from equal to unequal cluster sizes and thus from constant to varying inclusion probabilities, design weighting for the point estimator is necessary to get unbiased estimates. Weights, however, must be incorporated in the variance estimation as they introduce additional variance. If the variance estimator is used in the nominator of the design-based formula for the design effect, any increase in the

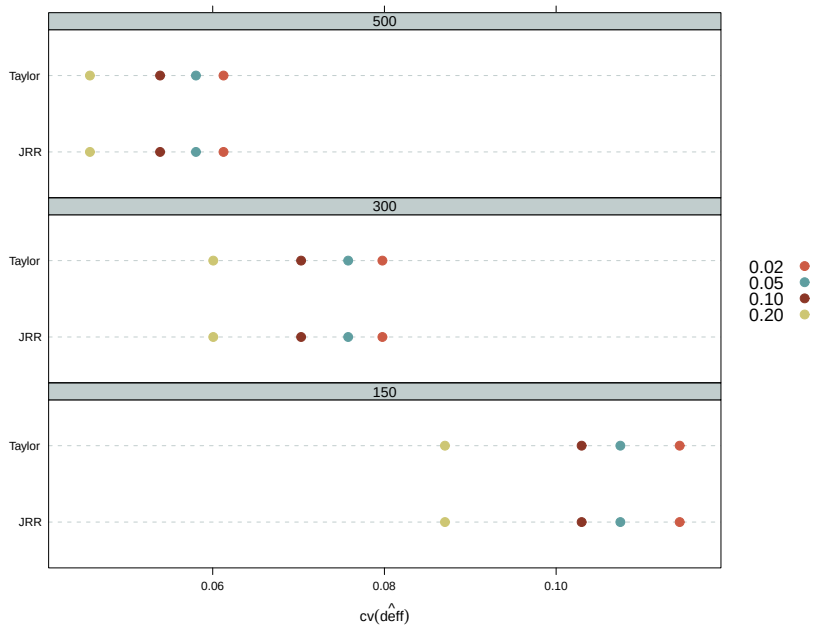


Figure 23: *Grouped dotplots of the cv of $\widehat{deff}$ under cluster sampling with unequal cluster sizes for given scenarios with continuous data*

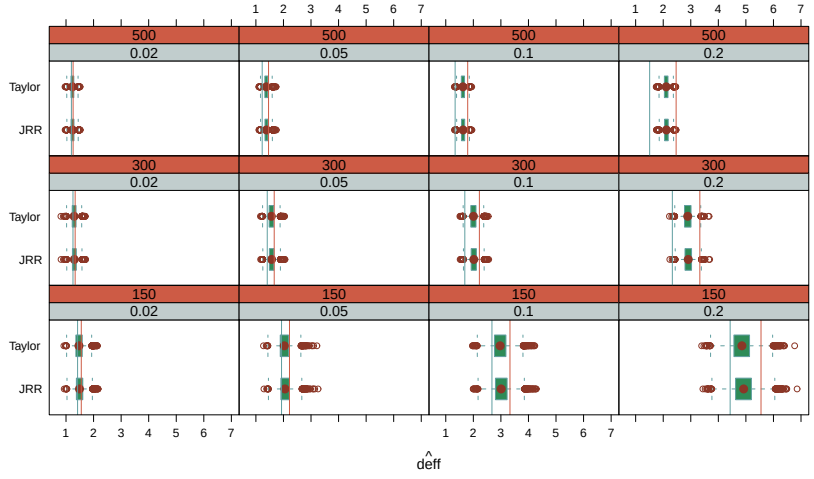


Figure 24: *Grouped boxplots of JRR and Taylor series estimates of $\widehat{deff}$ under cluster sampling with unequal cluster sizes for given scenarios with continuous data*

variance of the HT estimator will directly lead to an increase in the design effect as the denominator will not vary since weighting is completely ignored here. Figure 25 shows the mean estimated design effects in different scenarios with equal and unequal cluster sizes and hence with and without the effect of variation in design weights.

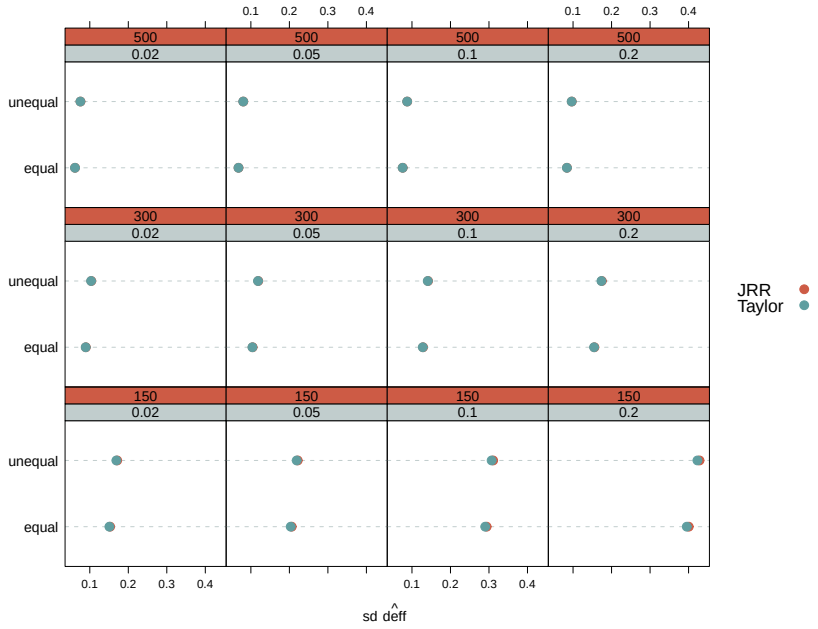


Figure 25: *Grouped dotplots of mean $\widehat{deff}$ under cluster sampling with equal and unequal cluster sizes for given scenarios with continuous data*

We can hardly see any differences between the JRR and the Taylor based estimator of *deff* as the plot symbols in all panels overlap almost perfectly. We do see, however, a difference in the magnitude of estimated *deff*. Both estimators reflect the additional variance introduced by variation in design weights as they point estimates show higher values with unequal cluster sizes (comparison of upper and lower part within a panel).

Similar patterns can be observed when we look at the precision of estimators. Figure 26 shows the coefficients of variation *cv* of the estimators.
Again, estimators are almost identical in terms of precision. Precision, however, increases with an increase in the population level of ρ and with a decrease in the average cluster size. At any given scenario the precision of both estimators is lower under cluster sampling with unequal than with equal cluster sizes.

5.4.1.4 Variance Estimation under Cluster Sampling assuming SRS

The design based estimators of the design effect make use of the data from the cluster sample at hand to estimate the denominator in equation (3.1) and (3.2), respectively. The precision of a design-based estimator of *deff* of course also depends on the quality of an appropriate variance estimator in the enumerator, but it will also be influenced by the degree to which the assumption of the cluster sample data to be independent

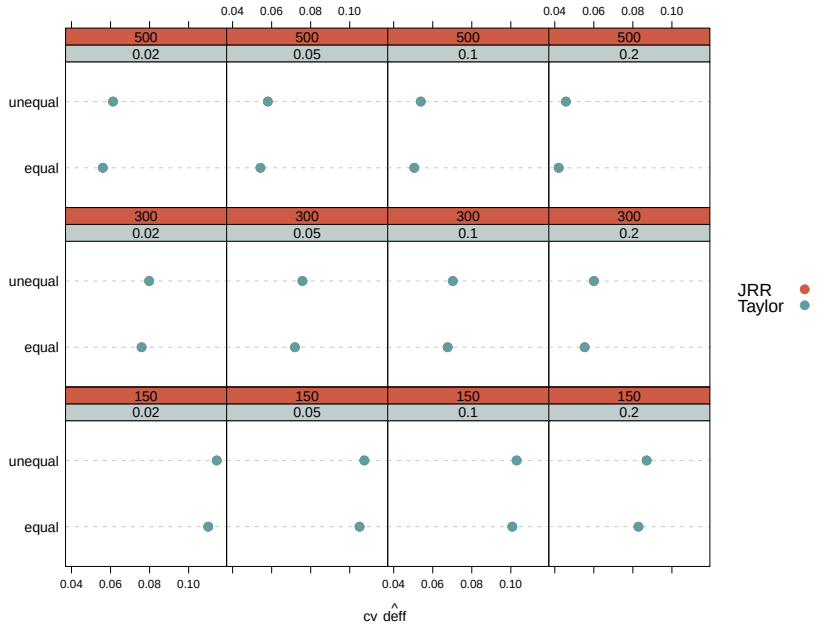


Figure 26: *Grouped dotplots of cv of $\widehat{deff}$ under cluster sampling with equal and unequal cluster sizes for given scenarios with continuous data*

identically distributed holds. The following simulation study investigates the patterns of the ratio

$$R = \frac{\widehat{Var}\left(\hat{\bar{y}}_{\text{clu2}}\right)}{\widehat{Var}\left(\hat{\bar{y}}_{\text{srs}}\right)} \quad , \tag{5.5}$$

where $\widehat{Var}\left(\hat{\bar{y}}_{\text{clu2}}\right) = \dfrac{Var\left(y_{\text{clu2}}\right)}{n_{\text{clu2}}}$ is the variance of the sample mean of a two-stage cluster sample of size n_{clu2} and $\widehat{Var}\left(\hat{\bar{y}}_{\text{srs}}\right) = \dfrac{Var\left(y_{\text{srs}}\right)}{n_{\text{srs}}}$ is the variance of the sample mean of a simple random sample of size n_{srs}.

Figure 27 shows the relative Root MSE (Rel. Root MSE) under different scenarios. Rel. Root MSE was calculated according to the following formula:

$$\text{Rel. Root MSE}\left(\hat{\theta}\right) = \sqrt{\frac{\sum_{i=1}^{5000}\left(\frac{\hat{\theta}_i - \theta}{\theta}\right)^2}{5000}},$$

with the mean of $\widehat{Var}\left(\hat{\bar{y}}_{\text{srs}}\right)$ over 5 000 iterations as true value θ and the estimated variances of $\hat{\bar{y}}$ based on cluster sample data, $\widehat{Var}\left(\hat{\bar{y}}_{\text{clu2}}\right)$, as $\hat{\theta}_i$. One can see that in all scenarios the Rel. Root MSE is larger when cluster sizes vary than in the case when they are equal. A direct comparison of Rel. Root MSE of $\hat{\theta} = \widehat{Var}\left(\hat{\bar{y}}_{\text{srs}}\right)$ and

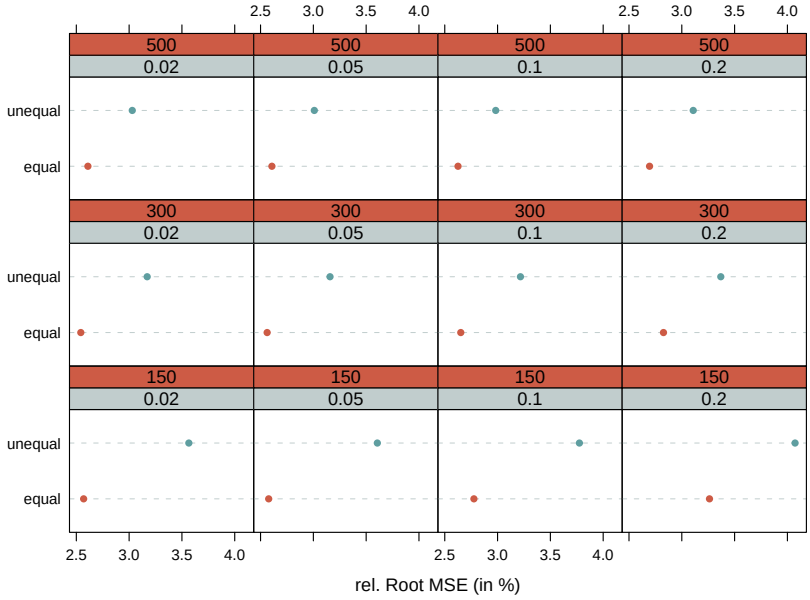


Figure 27: *Grouped dotplots of Rel. Root MSE of estimated variance of the sample mean under cluster sampling with continuous data*

$\hat{\theta} = \widehat{Var}(\hat{\bar{y}}_{\text{clu2}})$, respectively, is given in Figure 28. We can see that, little surprisingly,

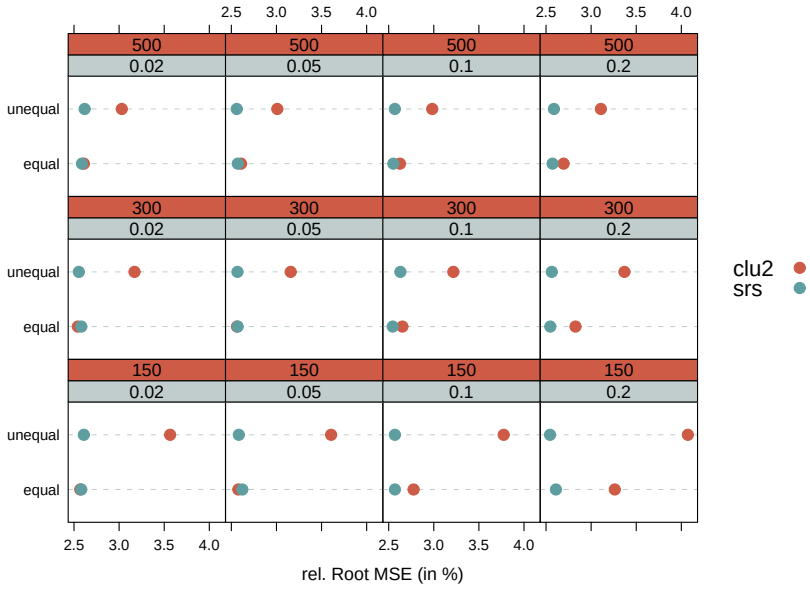


Figure 28: *Grouped dotplots of Rel. Root MSE of estimated variance of the sample mean based on clu2 and on real srs data with continuous data*

the Rel. Root MSE based on a srs is smaller than the respective Rel. Root MSE based on a clu2 in all scenarios. With an increase of the homogeneity in the population (comparison within a row) the Rel. Root MSE of the clu2 data increases. As (average) cluster sizes get smaller, the Rel. Root MSE based on clu2 data decreases (comparison within a column). When cluster sizes vary compared to the case when cluster sizes are equal (comparison within a panel), Rel. Root MSE is also bigger. This leads to the conclusion that the variance of the estimator based on the cluster sample at hand which is commonly used in the denominator of the formula for the design effect is, of course, less efficient than the estimator based on a srs of the (expected) same size. What is more interesting, however, is a comparison of relative bias between estimates based on clu2 and srs data. The following Figure gives an overview of the distribution of

$$\text{Rel. Bias}\left(\hat{\theta}\right) = \frac{\frac{1}{5\,000}\sum_{i=1}^{5\,000}\hat{\theta}_i - \theta}{\theta},$$

as before either for $\hat{\theta} = \widehat{Var}\left(\hat{\bar{y}}_{\text{srs}}\right)$ and $\hat{\theta} = \widehat{Var}\left(\hat{\bar{y}}_{\text{clu2}}\right)$, respectively. The figure shows

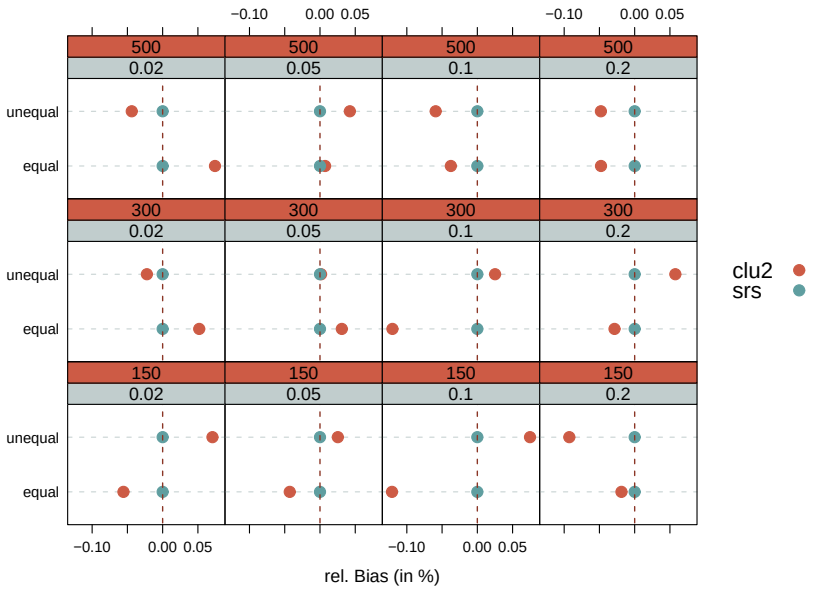


Figure 29: *Grouped dotplots of Rel. Bias of estimated variance of the sample mean based on clu2 and on real srs data with continuous data*

that, as expected, Rel. Bias $\left(\widehat{Var}\left(\hat{\bar{y}}_{\text{srs}}\right)\right)$ is effectively zero in all scenarios. Almost all Rel. Bias based on clu2 are, however, non-zero. Nevertheless, it has to be mentioned that the magnitude of bias is rather small, ranging from -0.1210% ($\rho = 0.10$, $m = 150$, equal cluster sizes) to 0.0747% ($\rho = 0.10$, $m = 150$, unequal cluster sizes). With a decrease in cluster sizes (comparison within a column), the estimates tend to be less downward biased in the scenario with equal cluster sizes (lower part of a panel)

and more downward biased with unequal cluster sizes (upper part of a panel). Both effects, however, are perfectly stringent. An increase in the population level of ρ (comparison within a row) tends to change the direction from upward to downward bias in scenarios with equal cluster sizes (again, lower part of a panel). This effect cannot be observed in scenarios with unequal cluster sizes. Here an increase in ρ in some cases increases and in other cases decreases Rel. Bias depending on both m and the levels of ρ for which the comparison is made. Nevertheless, one can state that $\widehat{Var}\left(\hat{\bar{y}}_{\text{srs}}\right)$ tends to be – at a low level – biased and will hence lead to biased estimates if naively used in the denominator of a design-based estimator of *deff*.

5.4.2 Estimation of the Design Effect for the Median

There exists no closed form of the model-based design effect for the median. Thus, estimation of *deff* for the median must be design-based. The JRR estimation technique can be used to construct a variance estimator of the median (see Section 3.3.1.2) with cluster sample designs. This estimator can then be used as the numerator in equation (3.2). We can use the variance estimator of the median assuming srs of elements proposed by McKean and Schrader (1984) as an appropriate denominator in equation (3.2). The variance estimator for the median by McKean and Schrader (1984) is defined as

$$\widehat{Var}_{\text{MS}}\left(\hat{\tilde{y}}\right) = \left\{\frac{y_{n-c+1} - y_c}{2(1.96)^2}\right\}^2 \quad , \tag{5.6}$$

where $c = \frac{n+1}{2} - 1.96\left(\frac{n}{4}\right)^2$. The ratio of these two quantities is calculated 10 000 times for each combination of ρ and m, i.e. the simulation set-ups follow the same logic as the ones described before. Figure 30 shows the distribution of the mean estimated design effect of the median based on the estimator described above. For a non-parametric measure like the median, the point estimate of $\widehat{deff}^{JRR}$ shows very similar patterns as in the case of the HT estimator for the population mean: With an increase of the homogeneity in the population, the estimated design effect increases and with small (average) cluster sizes it decreases. Also in line with the findings of the previous section is the observation that the design effect is larger if cluster sizes vary than if they are constant. Figure 31 shows the coefficient of variation of estimates of the $\widehat{deff}^{JRR}$ estimator for the median. We have to consider the coefficient of variation since the variance and the standard deviation of estimates will be influenced by the mean value of estimates which are, in turn influenced by levels of m and different levels of the population parameter ρ between which we want to make comparisons.

We can see that, standardized on the mean, the estimator of the design effect for the median is least precise when homogeneity in the population and (average) cluster sizes are small. Due to the fact that the expected value is larger with unequal cluster size, the denominator in the formula of the coefficient of variation rules the fraction, although the standard deviations of estimates under unequal cluster sizes are larger compared to the scenario with equal cluster sizes. Thus, for a comparison of the precision of the estimator between equal and unequal cluster sizes for given levels of

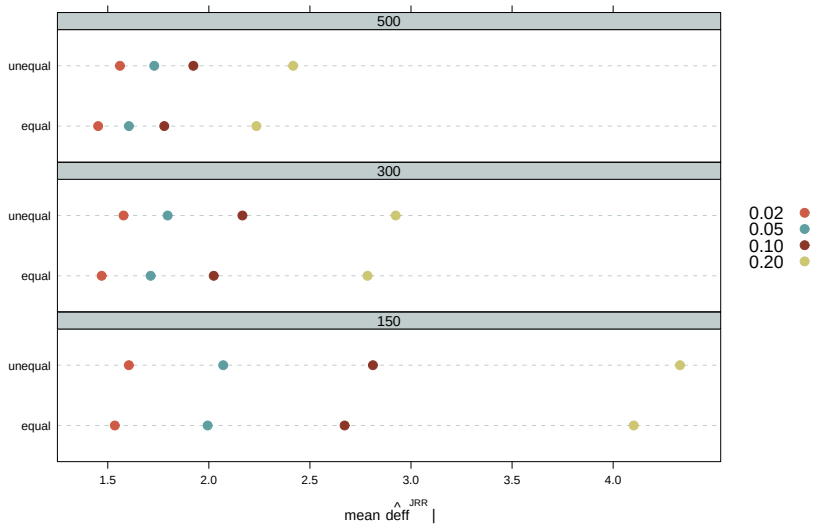


Figure 30: *Grouped dotplots of the mean estimated design effect of the median based on* $\widehat{deff}^{JRR}$ *under cluster sampling with equal and unequal cluster sizes for given scenarios with continuous data*

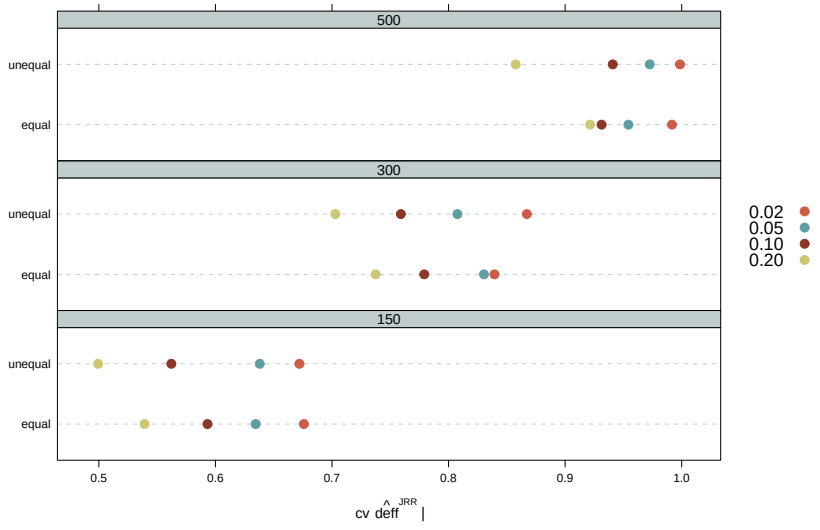


Figure 31: *Grouped dotplots of the coefficient of variation of the estimated design effect of the median based on* $\widehat{deff}^{JRR}$ *under cluster sampling with equal and unequal cluster sizes for given scenarios with continuous data*

m and ρ, we have to look at the standard deviation, not the coefficient of variation. This comparison reveals that in almost all cases the precision is higher in the scenario with equal than with unequal cluster sizes (see Figure 32).

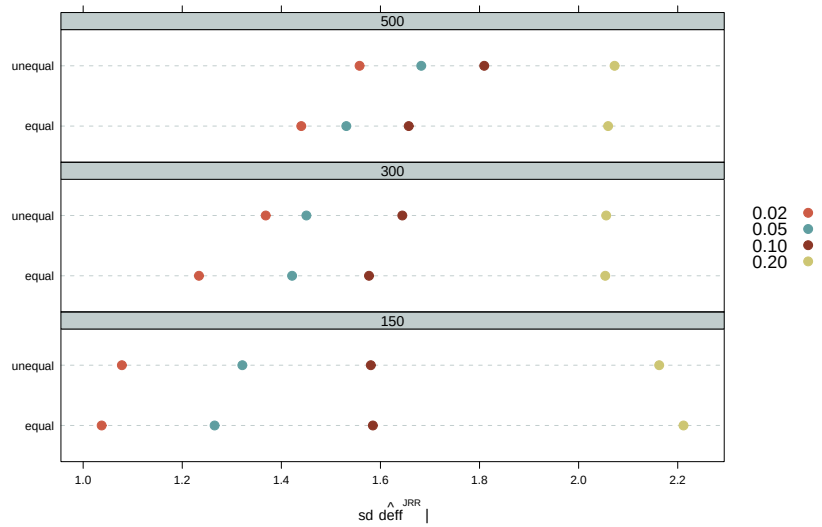


Figure 32: *Grouped dotplots of standard deviations of the estimated design effect of the median based on* $\widehat{deff}^{JRR}$ *under cluster sampling with equal and unequal cluster sizes for given scenarios with continuous data*

5.4.3 Dichotomous Data

The design-based variance estimators for the HT estimator used to estimate the design effect with binary data differ a bit from those used in the case of continuous data. The JRR and Taylor estimators of *deff* have to be modified to estimate the variance of an appropriate point estimator with binary outcome data (see Sections 3.3.1.1 and 3.3.1.2).

5.4.3.1 Cluster Sampling with equal Cluster Sizes

With binary outcome data, the additional parameter of π has to be considered when comparing the results of the estimated design effect. Hence, the following plots differentiate different levels of overall π by different colours. It can be seen from Figure 33 which depicts the cv of the estimates produced by the respective estimators in the scenarios of this setting. There are, however, hardly any differences bewtween estimators in a given scenario but huge differences between different levels of π within a given scenario. Depending on the scenario, $cv(\widehat{deff})$ is up to twice as large for $\pi = 0.05$ than for $\pi = 0.50$ (as for example in $\rho = 0.20$, $m = 150$). Put the other way around, the gain in efficiency is largest for a change in π from 0.05 to 0.25 and only marginal when π increases from 0.25 to 0.50. However, with small and medium population ρ there is hardly any difference in cv for different levels of π – especially for medium and small cluster sizes where $\widehat{deff}$ tends to be very small. In these scenarios, cv for $\pi = \{0.25, 0.50\}$ hardly varys with ρ. Nevertheless, the general tendency of the cv

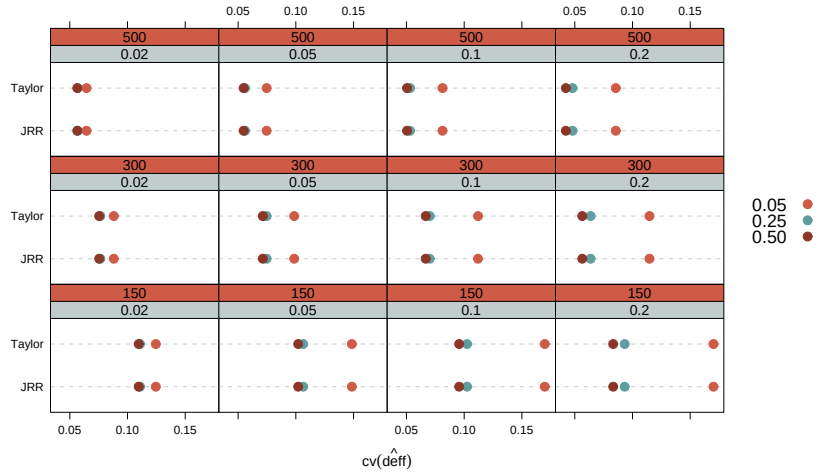


Figure 33: *Grouped dotplots of the cv of $\widehat{deff}$ under cluster sampling with equal cluster sizes for given scenarios with binary data*

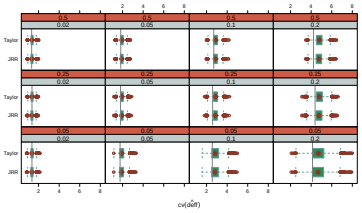
for JRR and Taylor estimator is to show less variation when ρ is large and π is not too small.

When we take a look at the full distribution of estimates in Figure 34, we can see that also with binary data, both estimators have positive bias for the Monte Carlo estimated true design effect. The JRR estimator, however, shows a little less bias than the Taylor estimator. The magnitude of bias is only slightly influenced by π as the boxplots' location hardly changes when moving vertically through the lattice plot. It is more depending on population ρ since the boxplots' location moves further away from the grey line when moving through the lattice from low to high values of ρ.

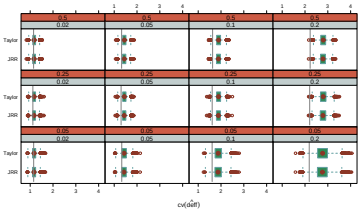
5.4.3.2 Cluster Sampling with unequal Cluster Sizes

Under cluster sampling with unequal cluster sizes (and unequal inclusion probabilities), the additional variation introduced by weighting is reflected in increased *cv* of the estimates produced by the simulation runs. Figure 35 gives an overview over the distribution of the estimators' *cv* under different scenarios. There are, however, no changes in the interrelation of the estimators as compared to the previous setting. That is, also when using design weights the JRR estimator in every scenario has smaller standard deviation than the Taylor estimator. But since JRR's mean over 10 000 iterations is also smaller, it has essentially the same *cv* as the Taylor estimator.

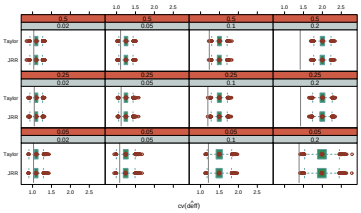
These tendencies are both graphically illustrated by the grouped boxplots of Figure 36. Here we can observe a) the biased nature of both estimators, b) the larger variance of the Taylor estimator and c) the smaller average of the JRR estimator. As a rule, the JRR is less biased than the Taylor estimator. The bias is smaller for small values of ρ and large average cluster sizes. The magnitude of π seems to have only a very small effect on bias.



(a) $m = 150, \bar{b} = 20$



(b) $m = 300, \bar{b} = 10$



(c) $m = 500, \bar{b} = 6$

Figure 34: *Grouped boxplots of JRR and Taylor series estimates of $\widehat{deff}$ under cluster sampling with equal cluster sizes for given scenarios with binary data*

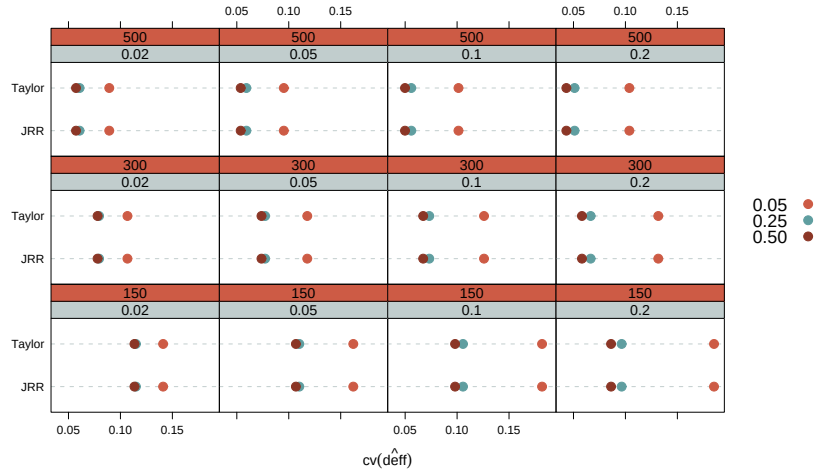
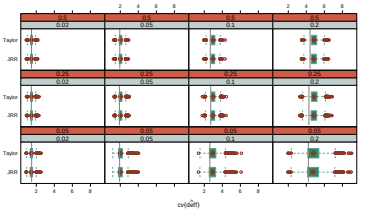


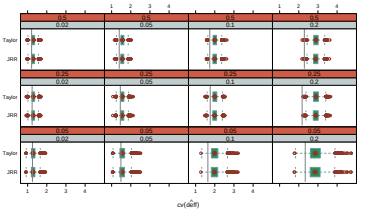
Figure 35: *Grouped dotplots of the cv of $\widehat{deff}$ under cluster sampling with unequal cluster sizes for given scenarios with binary data*

5.4.3.3 Comparison of Two-Stage Cluster Sampling with equal and unequal Cluster Sizes

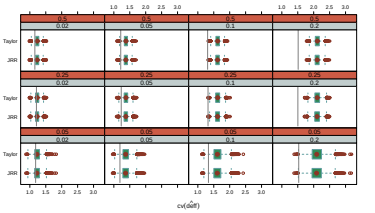
The loss in efficiency as measured by cv can be severe in a setting with unequal cluster sizes compared to the same setting with equal cluster sizes. Figure 37 displays



(a) $m = 150, \bar{b} = 20$



(b) $m = 300, \bar{b} = 10$



(c) $m = 500, \bar{b} = 6$

Figure 36: *Grouped boxplots of JRR and Taylor series estimates of $\widehat{deff}$ under cluster sampling with unequal cluster sizes for given scenarios with binary data*

the ratio $R_{cv} = \dfrac{cv\left(\widehat{deff}_{(+)[\text{un}]\langle+\rangle\{+\}}\right)}{cv\left(\widehat{deff}_{(+)[\text{eq}]\langle+\rangle\{+\}}\right)}$. Here we can see that this ratio can be as big as 1.4 indicating a 40% increase in the coefficient of variation when weighting is applied as to the setting where all weights are constant. What can be seen from the above figure is that as π gets very small, this has an enormous effect on the magnitude of R_{cv}. That is, within a given scenario a change from $\pi = 0.50$ to 0.25 has hardly any effect on R_{cv}. A change from $\pi = .25$ to $\pi = 0.05$, however, leads to a dramatic increase in R_{cv}. This effect is most significant when, in addition, ρ is small and m is large (and hence the average cluster size is small).

If we now take a look at the ratio of the averages of estimated design effect, $R_{mean} = \dfrac{\overline{\widehat{deff}_{(+)[\text{un}]\langle+\rangle\{+\}}}}{\overline{\widehat{deff}_{(+)[\text{eq}]\langle+\rangle\{+\}}}}$, we can observe different effects. Figure 38 shows R_{mean} for different scenarios. First of all, there are almost no differences between estimators under study. Furthermore, the magnitude of π has hardly any influence on the average of the estimated design effect. A significant effect, however, on the magnitude of R_{mean} is introduced both by average cluster size and by the magnitude of the population parameter ρ. With decreasing cluster size and a decrease in ρ, R_{mean} increases. This indicates that the effect of additional variance introduced by weights is more pronounced when the effect of clustering is rather small.

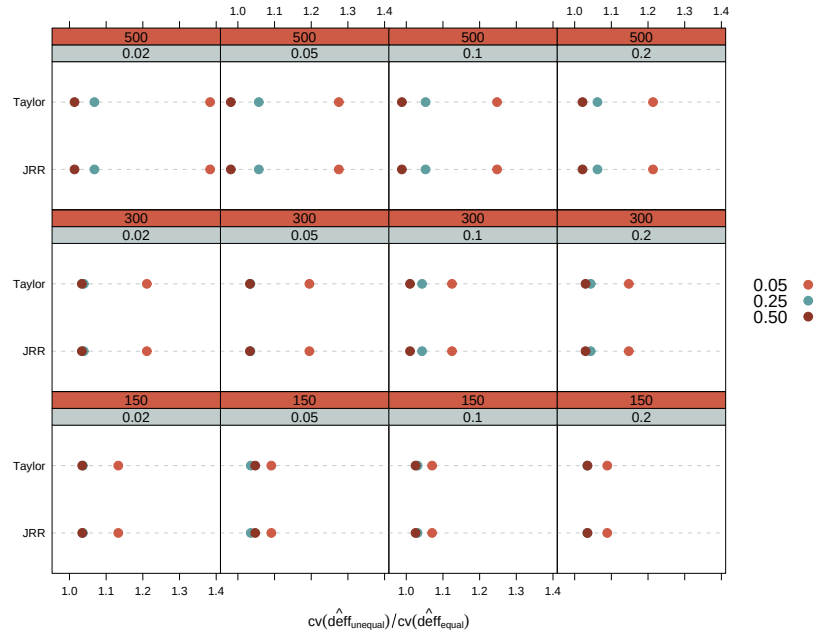


Figure 37: *Grouped dotplots of the ratio of cvs of cluster sampling with unequal to equal cluster sizes for given scenarios with binary data*

5.5 Model-based Estimation of the Design Effect

Estimation of the design effect in a model-based manner requires the estimation and/or calculation of the two components of the design effect, namely $\widehat{deff}_c$ and $\widehat{deff}_p$. These quantities, in turn, depend on precise estimation of the components which they are a function of. For the estimator $\widehat{deff}_c$, a crucial task is the estimation of ρ and calculation of b^*, while to estimate $\widehat{deff}_p$, only design weights have to be calculated. Since this task, as well as the calculation of b^*, is rather trivial, the following Monte Carlo studies are limited to an investigation of the quality of estimators of ρ.

Due to the fact that, depending on the scale, different estimators for ρ have been proposed, the simulation studies consider *Gaussian* as well as *binary* outcome data separately.

The Gaussian setting is similar in set-up to the situation in the previous section, thus leading to a 2 (cluster size type) × 3 (average cluster size) × 8 (values of ρ)=48 factorial study design. The binary setting has 2 (cluster size type) × 3 (average cluster size) × 8 (values of ρ) × 3 (values of π)=144 factors, which is also similar to the setting described in section 5.4. In the scenarios of the first setting, the whole set of estimators described in section 4.2 is investigated. In the setting with binary outcome data, nine estimators – described in section 4.3 – are evaluated. In the following subsections, first the behaviour of the estimators of ρ is illustrated for continuous (section 5.5.1) and, second, for binary (section 5.5.2) data.

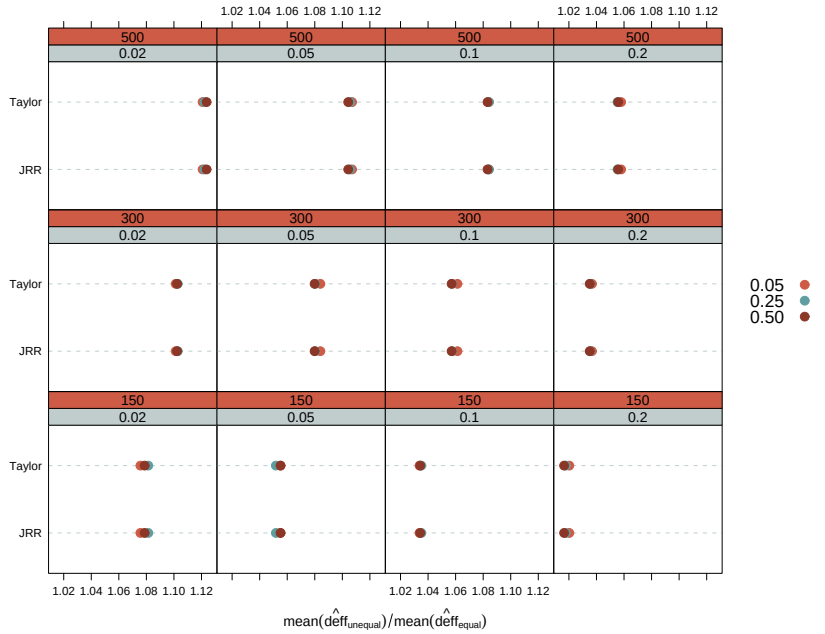


Figure 38: *Grouped dotplots of the ratio of averages of cluster sampling with unequal to equal cluster sizes for given scenarios with binary data*

5.5.1 Estimation of ρ with Continuous Data

A large number of estimators for the intraclass correlation coefficient for continuous outcome data has been proposed in the literature. An overview is given in Section 4.2. These estimators behave differently in different environments. Some estimators react sensitive to variations in cluster size whereas others tend to underestimate the parameter of interest when cluster sizes vary. A first natural distinction between estimators of ρ is made in Section 4.2 where the classical ANOVA and F-Type estimators form one group and estimators based on the variance decomposition of a random effects model form the second group. Additionally, the effect of equal and unequal cluster sizes is of interest in the evaluation of the estimators' quality[16]. Finally, the magnitude of bias and precision also depends on the magnitude of the population parameter, ρ. Thus, also ρ is considered as a factor in the Monte Carlo simulation and in the analysis.

5.5.1.1 Two-Stage Cluster Sampling with equal Cluster Sizes

First, the case of cluster sampling with equal cluster sizes (i.e. equal inclusion probabilities) is considered. In this scenario, where all sampled clusters have equal size, $\hat{\rho}^{(\mathrm{AOV})}$, $\hat{\rho}^{(\mathrm{F2})}$ and $\hat{\rho}^{(\mathrm{REML})}$ are equivalent as well as $\hat{\rho}^{(\mathrm{FR})}$ and $\hat{\rho}^{(\mathrm{ML})}$; when ignoring

16 Please note that also with unequal cluster sizes no weighting has been applied since these estimators are not designed to take weights into account.

circumstances where classical estimators yield negative estimates and random effects model based estimators yield zero. The grouped boxplots of Figure 39 give an overview of the distribution of the estimators in given scenarios. Within each panel,

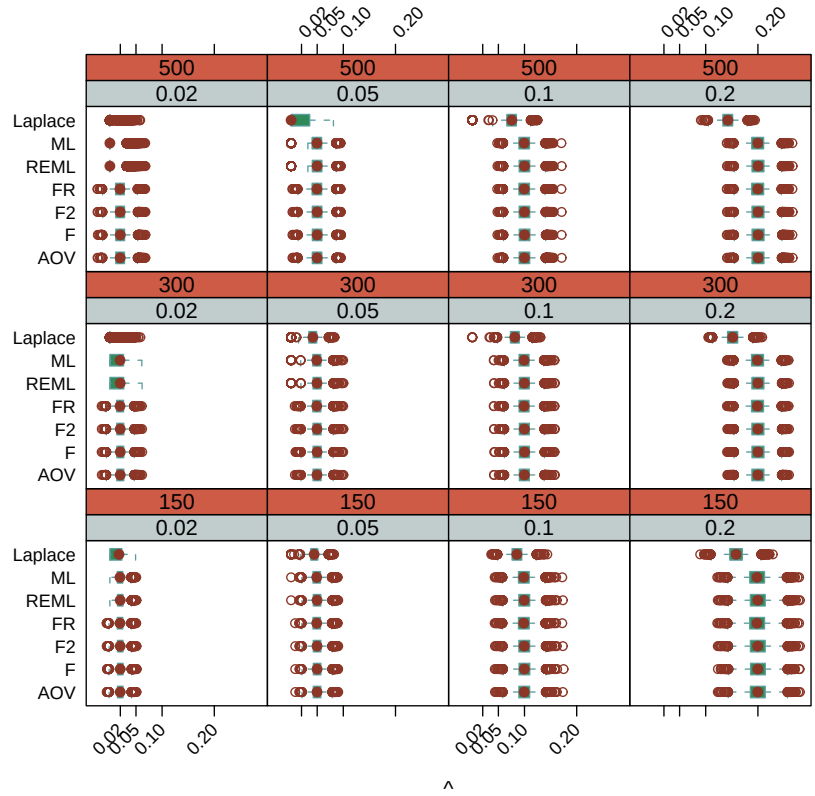
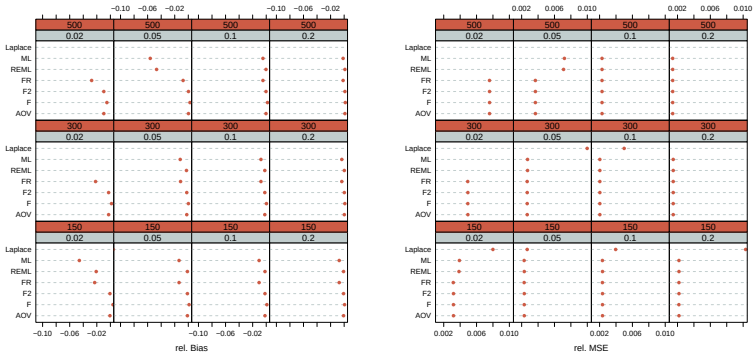


Figure 39: *Grouped boxplots of estimates of ρ under two-stage cluster sampling with equal cluster sizes for continuous data*

a comparison between estimators can be made for given levels of population ρ and cluster size. The distribution of estimates for different levels of ρ in the population shows larger variation for high levels of ρ. All estimators react in similar manner to an increase in ρ at a given cluster size. Especially in the case of $\rho = 0.02$, one can clearly observe that the REML, ML and the Laplace estimators are restricted in range to a minimum of zero. In addition, we can observe a clear tendency of the Laplace estimator to yield downwards biased estimates.

Bias and precision – in terms of relative mean squared error (Rel. MSE) – of the estimators are illustrated in more detail by the grouped dotplots of Figure 40. As already indicated by Figure 39, the classical estimators show hardly any bias and are very precise; also in the case of very low population ρ and very small cluster sizes, the only exception being the FR estimator which shows some downward bias



(a) relative Bias (b) relative Mean Squared Error

Figure 40: *Grouped dotplots of Rel. Bias and Rel. MSE of estimators of ρ under two-stage cluster sampling with equal cluster sizes for continuous data*

in all scenarios. The estimators of the random effects model class are, in contrast, all biased – some of them in certain scenarios quite heavily. The relative bias of the Laplace estimator has already been mentioned before and becomes obvious also in Figure 40, as it reaches a downward bias of up to -87.5% and even under the modest conditions it is at an unacceptable level of -20.8% which is, due to the choice of scale of the x-axis, not even shown in the plots[17]. The REML and ML estimators are less biased with the REML estimator showing less bias than the ML estimator. However, REML (as well as ML and especially Laplace) estimator tends to react very sensitive on a decrease of cluster size, both in terms of bias and precision – especially when population ρ is small[18]. The least biased estimator under each scenario is given in the cells of Table 5. The most successful estimator in this respect is the F estimator which is the least biased one in seven out of twelve scenarios. It is followed by F2 and REML which are the least biased estimators in three and two scenarios, respectively. These differences cannot, however, be assured by usual significance tests. Turning to

Table 5: *Least absolute bias of all estimators by population ρ and average cluster size*

	0.02	0.05	0.10	0.20
150	F2	F	F	F
300	F	F	F	F2
500	F	F2	REML	REML

precision as measured by Rel. MSE, the classical as well as the random effects model estimators react sensitively to a decrease in cluster size – again, no more so than for low levels of population ρ. For a given level of ρ and a given cluster size, one can

17 For better comparison of the estimators the scale of the x-axis was chosen very narrow.

18 In fact, REML and ML are both so severly biased that they are outside the scale of the x-axis in the two most extreme settings.

hardly see any difference between estimators except for an unacceptable high Rel. MSE of the REML, ML and Laplace estimators in extreme scenarios and of the Laplace estimator in all scenarios. However, with moderate levels of ρ and small to medium cluster sizes, REML and ML estimators have Rel. MSE comparable the the classical estimators. As can be seen from the following table, the FR estimator is most precise in eight out of twelve scenarios, F and ML estimators only twice each. An interesting

Table 6: *Least Rel. MSE of all estimators by population ρ and average cluster size*

	0.02	0.05	0.10	0.20
150	FR	FR	FR	F
300	FR	FR	ML	FR
500	FR	FR	F	ML

feature of the joint effects of cluster size and population ρ on Rel. MSE is that with $\rho = \{0.02, 0.05\}$ a decrease in cluster size increases Rel. MSE. With high levels of ρ (i.e. 0.10 and 0.20), however, the reverse effect can be observed – a decrease in cluster size now leads to a decrease in Rel. MSE. Table 23 in the appendix summarizes Rel. Bias and Rel. MSE of the estimators under given simulation scenarios.

5.5.1.2 Two-Stage Cluster Sampling with unequal Cluster Sizes

When allowing for variations in cluster sizes, the picture that emerged in the previous subsection changes a bit. However, the basic underlying structure stays the same, as Figure 41 indicates. In extreme scenarios (small value of the parameter ρ, small average cluster sizes), all estimators have difficulties yielding unbiased and precise estimates. If we take a closer look, however, we can now observe variation between estimators of all types – both in terms of bias and precision. These differences are most obvious in extreme scenarios. Although the classical estimators, again, show very little bias also in extreme scenarios, the F estimator can be seen to have relatively high variation compared to all other classical estimators.

Looking at Figure 42, we can observe a similar pattern as in the case of equal cluster sizes. A certain degree of bias can now be observed also for the FR estimator. The REML, ML and Laplace estimators are severely biased in extreme scenarios[19]. As a rule, a decrease of average cluster size at a given level of ρ leads to an increase in the magnitude of relative bias. This effect, however, is only consistent when population ρ is small and for the random effects model estimators as well as for $\hat{\rho}^{(\mathrm{AOV})}$ and $\hat{\rho}^{(\mathrm{F})}$. For higher levels of ρ, estimators seem to be less biased for small average cluster sizes at first glance. In fact, however, most of the estimators (especially the ones based on random effects models), simply change sign from negative to positive bias.

Table 7 indicates that the F estimator shows least absolute bias in six of the 12 scenarios, AOV, REML and F2 in two settings each. With big and medium average cluster sizes and small population ρ, the F estimator can be seen to be very unprecise

19 Again, they are so heavily biased that they lay outside the x-axis scale.

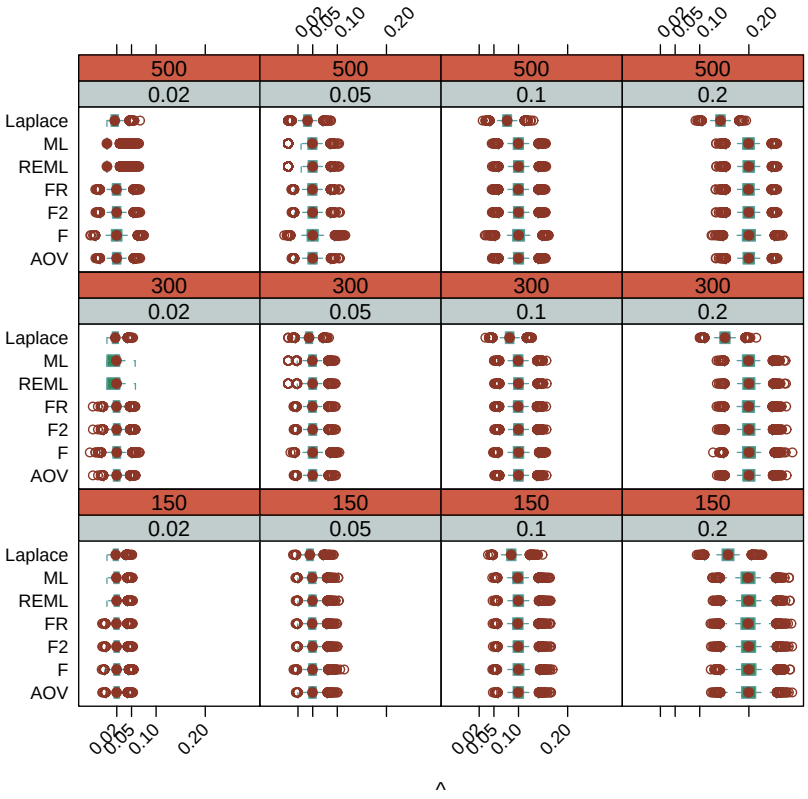
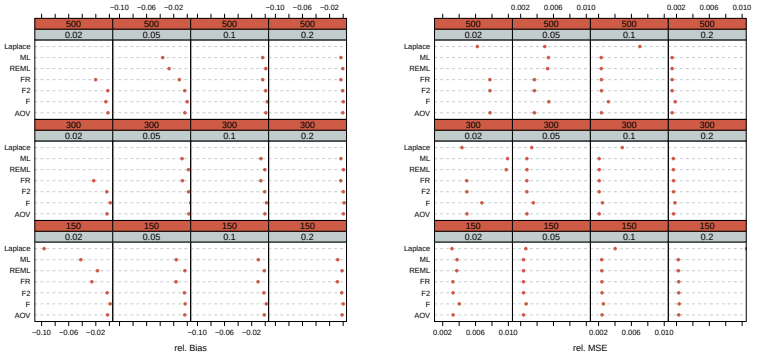


Figure 41: *Grouped boxplots of estimates of ρ under two-stage cluster sampling with unequal cluster sizes for continuous data*



(a) relative Bias (b) relative Mean Squared Error

Figure 42: *Grouped dotplots of Rel. Bias and Rel. MSE of estimators of ρ under two-stage cluster sampling with unequal cluster sizes for continuous data*

Table 7: *Least absolute bias of all estimators by population ρ and average cluster size*

	0.02	0.05	0.10	0.20
150	F	F	REML	F
300	F	REML	AOV	F2
500	AOV	F	F2	F

– in almost all scenarios it is the second most unprecise estimator after $\hat{\rho}^{(\text{Laplace})}$. Also with unequal cluster sizes the effect of an increase of Rel. MSE for $\rho = \{0.02, 0.05\}$ and a decrease for high values of ρ can be observed for almost all estimators. When it comes to precision, the ML estimator is most precise in five scenarios, FR and Laplace in three scenarios and REML in one as can be seen from Table 8 An overview of

Table 8: *Least Rel. MSE of all estimators by population ρ and average cluster size*

	0.02	0.05	0.10	0.20
150	Laplace	ML	ML	ML
300	Laplace	FR	FR	ML
500	Laplace	FR	ML	REML

relative bias and MSE for the complete set of scenarios is given in Table 24 in the appendix.

5.5.1.3 Comparison of Cluster Sampling with equal and unequal Cluster Sizes

In practical survey sampling one is of course only very rarely faced with a cluster sample that has PSUs of equal size. Nevertheless, many estimators (in fact all the classical ones) are based on that assumption. We have seen that in neither of the settings described before, the same estimator is both least biased and most precise at a given scenario. This makes the choice of an optimal estimator for a given complex sample even more complicated. Another way, hence, to evaluate the quality of an estimator is to look at its sensitivity to violations of the equal cluster size assumption.

The following figures depict the ratio of Rel. Bias and Rel. MSE for each estimator in each scenario of the unequal and the equal cluster size setting, hence the ratios

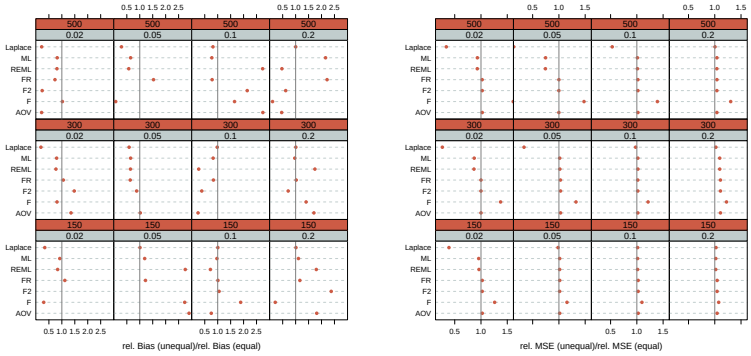
$$R^{\hat{\rho}^{(+)}}_{\text{Rel. MSE}} = \frac{\text{Rel. MSE}\left(\hat{\rho}^{(+)}_{(\text{clu2})\,[\text{ue}]\,\langle+\rangle\,\{+\}}\right)}{\text{Rel. MSE}\left(\hat{\rho}^{(+)}_{(\text{clu2})\,[\text{eq}]\,\langle+\rangle\,\{+\}}\right)}$$

and

$$R^{\hat{\rho}^{(+)}}_{\text{Rel. Bias}} = \frac{\text{Rel. Bias}\left(\hat{\rho}^{(+)}_{(\text{clu2})\,[\text{ue}]\,\langle+\rangle\,\{+\}}\right)}{\text{Rel. Bias}\left(\hat{\rho}^{(+)}_{(\text{clu2})\,[\text{eq}]\,\langle+\rangle\,\{+\}}\right)}.$$

To avoid confusion, Rel. Bias is treated unsigned. The x-axis of the Rel. Bias plots is

truncated to 3.0 but covers the full range toward the minimum. Thus, ratios of Rel. Bias not shown in a plot are above 3.0. The classical estimators in many scenarios are



(a) relative Bias (unsigned)

(b) relative Mean Squared Error

Figure 43: *Grouped dotplots of Rel. Bias and Rel. MSE of estimators of ρ under two-stage cluster sampling for unequal to equal cluster sizes for continuous data*

more biased (i.e. the ratio of unsigned Rel. Biases exceeds one) when cluster sizes vary compared to the case of equal cluster sizes. The ML, REML and Laplace estimators tend to react with an increase in relative downward bias on a decrease in cluster size at any level of ρ. When precision of estimators is concerned, only $\hat{\rho}^{(\mathrm{F})}$ and $\hat{\rho}^{(\mathrm{Laplace})}$ react sensitive on a change from equal to unequal cluster sizes. $\hat{\rho}^{(\mathrm{F})}$ tends to be less precise in a scenario with unequal cluster sizes. This loss in precision increases with a decrease in cluster size and a decrease in magnitude of the population parameter of ρ. The $\hat{\rho}^{(\mathrm{Laplace})}$ estimator, on the other hand, is more efficient with unequal cluster sizes than with equal cluster sizes – especially for small values of ρ in the population and small cluster sizes. This finding is a bit odd and counter-intuitive but may be an artifact which can be explained by the overall low level of precision of the Laplace estimator.

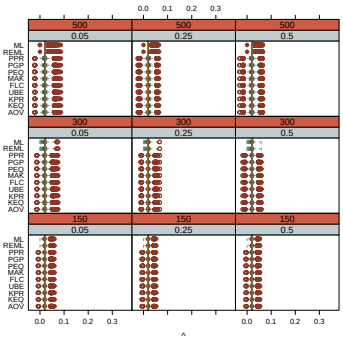
5.5.2 Estimation of ρ with Binary Data

The precision and Bias of estimators of ρ in the case of binary study variables is analysed in the following subsections. As mentioned earlier, also for binary outcome data, a wide range of estimators have been proposed in the literature. In this simulation study, however, only nine classical estimators and three estimators based on a variance decomposition of a random effects model are considered.

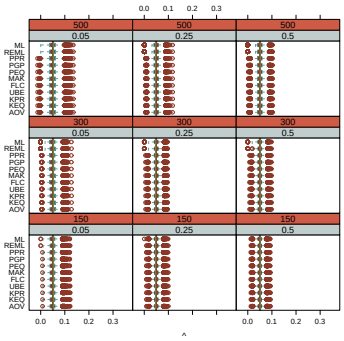
5.5.2.1 Two-Stage Cluster Sampling with equal Cluster Sizes

To get a first impression, the distribution of the estimates based on 10 000 repeated samples from the respective population is summarized in the grouped boxplots of Figure 44. The grouping is by estimator type, number of clusters and π for selected

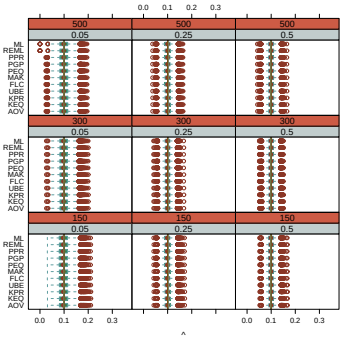
levels of ρ. Within each panel, hardly any difference between estimators occurs. This



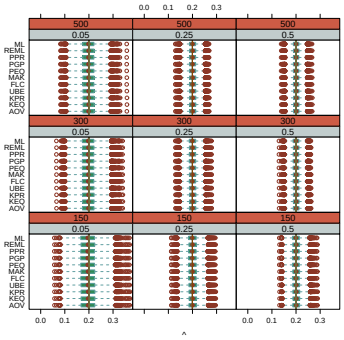
(a) $\rho = 0.02$



(b) $\rho = 0.05$



(c) $\rho = 0.10$

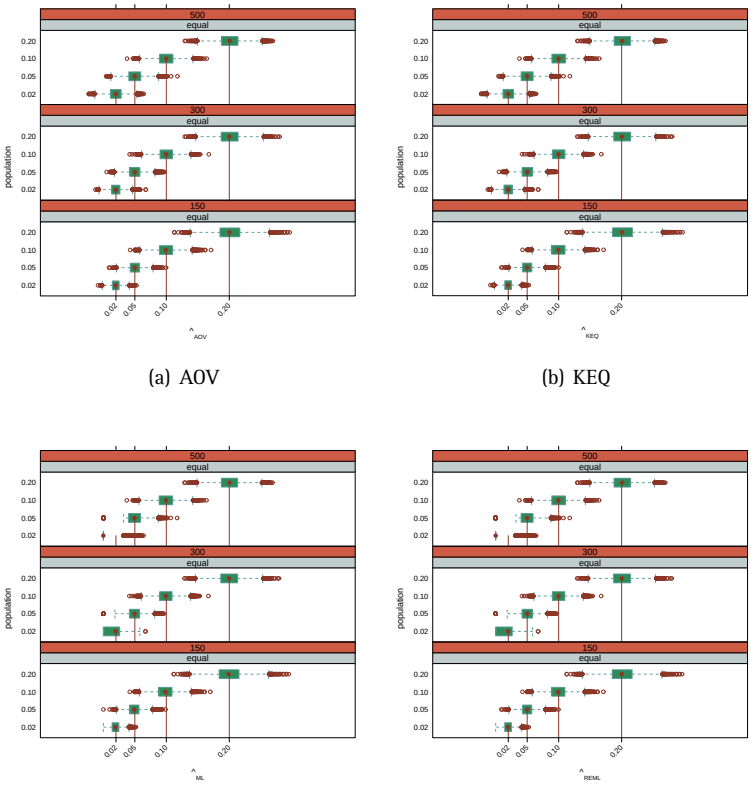


(d) $\rho = 0.20$

Figure 44: *Grouped boxplots of estimates of ρ under two-stage cluster sampling with equal cluster sizes for binary data*

is in line with what Mak (1988) found. We can, however, observe a pattern between panels within a plot at a given level of ρ. With an increase in the number of clusters (and hence a decrease in cluster size), estimators tend to be less precise. However, with a decrease in overall π, estimators tend to be more precise. These effects can also be observed in interaction. They are more obvious for large values of ρ (e.g. 0.20) than for small ones (e.g. 0.02).

Let us take a closer look at a set of selected estimators. I chose $\hat{\rho}^{(\text{AOV})}$, $\hat{\rho}^{(\text{KEQ})}$ and $\hat{\rho}^{(\text{ML})}$ as well as $\hat{\rho}^{(\text{REML})}$ estimators to illustrate a) that there is hardly any difference between the two typical estimators of the classical type and b) the differences in range between the classical estimators and the ones based on a variance decomposition of a random effects model. The plots of Figure 45 represent the simulation subset with $\pi = .25$ for usual levels of ρ. The AOV and the KEQ estimator both tend to slight overestimation especially with high levels of population ρ. For population $\rho = 0.02$,



(a) AOV (b) KEQ

(c) ML (d) REML

Figure 45: *Grouped boxplots of selected estimators of ρ under two-stage cluster sampling with equal cluster sizes for binary data; $\pi = 0.25$*

it is easy to see that the two other estimators' range is bounded to 0. Otherwise, all estimators behave very much the same, there is no severe deviation from the population value: AOV, KEQ and REML estimators are equivalent for the special case of equal cluster sizes (Mak, 1988).

The variation of the estimators depends on the magnitude of the population parameter to estimate. This is why the standard deviation as a measure of the spread of the distribution of estimates alone is misleading. Its interpretation must be accompanied by the relative Root Mean Squared Error (rRMSE) which is graphically illustrated in Figure 46. It must be noted that the x-axis' range is restricted and hence values outside that range will not be displayed. Again, we see that especially for small values of ρ and small cluster sizes, the rRMSE is rather big for all estimators – in fact larger than 0.5 for the combination $\rho = .02$ and $m = 500$. A rather odd pattern must be mentioned, too: rRMSE tends to be smaller for small cluster

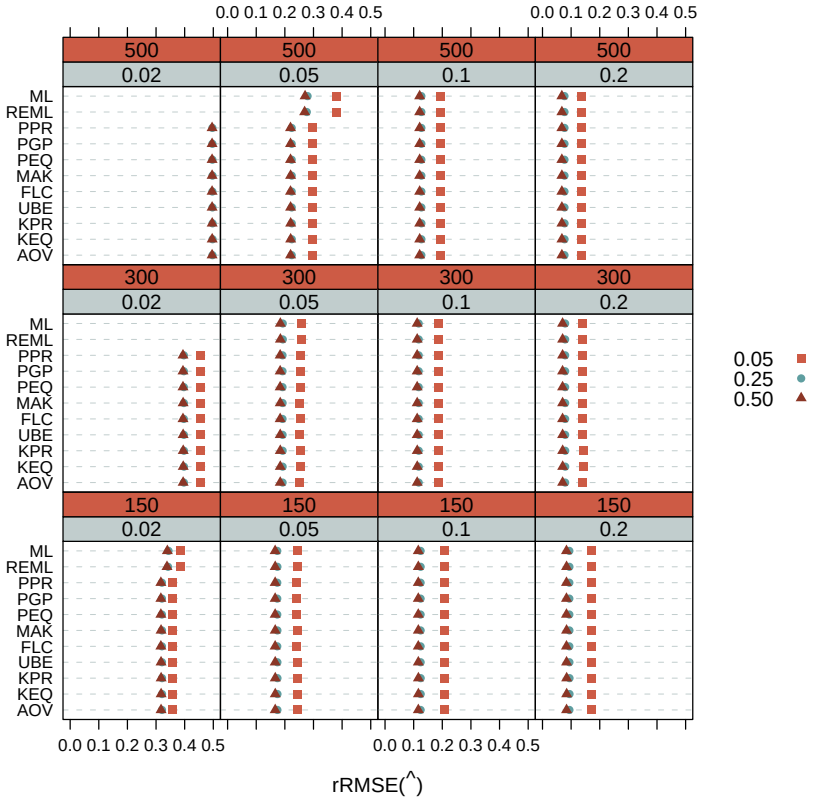
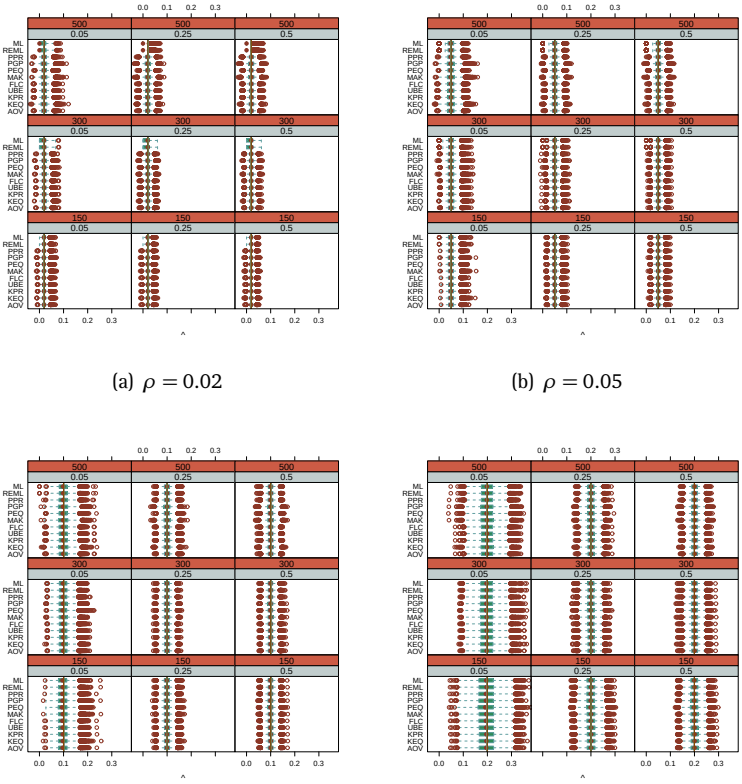


Figure 46: *Dotplot of rRMSE of estimators of ρ based on cluster sampling with equal cluster sizes*

sizes than for large ones for large values of ρ – especially for small values of π. If we look at the coefficient of variation, we can see that this is mainly due to a decrease in the variation of the estimators with smaller average cluster sizes as the share $\frac{mean\left(\hat{\rho}^{(AOV)}_{(clu2)[eq]\langle 150\rangle\{.20\}}\right)}{mean\left(\hat{\rho}^{(AOV)}_{(clu2)[eq]\langle 500\rangle\{.20\}}\right)} = .99$ but $\frac{sd\left(\hat{\rho}^{(AOV)}_{(clu2)[eq]\langle 150\rangle\{.20\}}\right)}{sd\left(\hat{\rho}^{(AOV)}_{(clu2)[eq]\langle 500\rangle\{.20\}}\right)} = 1.26$ thus ruling the 27% increase in $cv\left(\hat{\rho}^{(AOV)}_{(clu2)[eq]\langle 150\rangle\{.20\}}\right) = 0.2203$ to $cv\left(\hat{\rho}^{(AOV)}_{(clu2)[eq]\langle 500\rangle\{.20\}}\right) = 0.1731$.

5.5.2.2 Two-Stage Cluster Sampling with unequal Cluster Sizes

Things change when cluster sizes are allowed to vary. There are, however, the same trends of estimators to be more precise with large π, average cluster sizes and with large values of ρ. In this setting, however, we can also observe differences between estimators as they react at different magnitude to variations in cluster sizes. Figure 47 depicts the distribution of estimators by m and ρ for given levels of π. With unequal



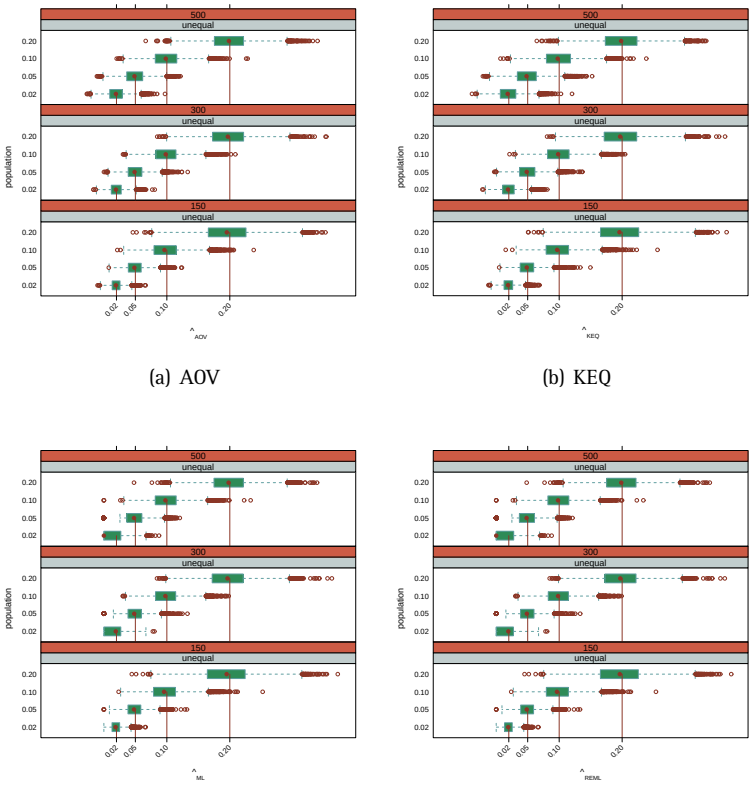
(a) $\rho = 0.02$ (b) $\rho = 0.05$

(c) $\rho = 0.10$ (d) $\rho = 0.20$

Figure 47: *Grouped boxplots of estimates of ρ under two-stage cluster sampling with unequal cluster sizes for binary data*

cluster sizes, we can observe differences between the distributions of estimators within each panel. A wide spread boxplot with far outliers indicates that an estimator is less precise than an estimator with a narrower box and less severe outliers. The KEQ, MAK and PGP estimators, for example, tend to show this pattern – especially for small values of π and small average cluster sizes.

If we take a closer look at selected estimators again, we can now observe differences and patterns of gain or loss in their precision. Figure 48 firstly illustrates the differences in the lower bounds of the classical (here: AOV and KEQ) estimators compared to the estimators based on the variance components of a random effect model (here: random effect model estimation based on ML and REML technique). The lower whisker of the boxplots (or even the inner fence) in Figures 5.48(c) and 5.48(d) ob page 102 never go beyond zero which leads to skewed distributions of $\hat{\rho}^{(\mathrm{ML})}$ and $\hat{\rho}^{(\mathrm{REML})}$ – this is most obvious in scenarios where the population parameter to estimate is rather small (i.e. ρ=0.02 or 0.05). A common pattern of all estimators presented here is the



(a) AOV (b) KEQ

(c) ML (d) REML

Figure 48: *Grouped boxplots of selected estimators of ρ under two-stage cluster sampling with unequal cluster sizes for binary data; $\pi = .25$*

loss of precision when cluster sizes decrease. This effect is amplified also when ρ in the population is rather small.

5.5.2.3 Comparison of Cluster Sampling with equal and unequal Cluster Sizes

A comparison between cluster sampling with equal and unequal cluster sizes must consider both, bias and precision. Figures 49 and 50 are very similar. They show the the ratio of means (Figure 49) and standard deviations (Figure 50) of estimators of ρ for levels of overall π. Switching from equal to unequal cluster sizes has the least effect on the bias of an estimator if $\pi = 0.50$ as in all panels of Figure 49 the ratio of means of almost all estimators in a majority of settings is closest to one if the overall success rate of study variable is 50% (indicated by a red triangle). Another pattern that holds for almost all estimators and settings is that bias decreases with large values of population ρ. The $\hat{\rho}^{(\mathrm{ML})}$ and $\hat{\rho}^{(\mathrm{REML})}$ estimators tend to react very sensitive

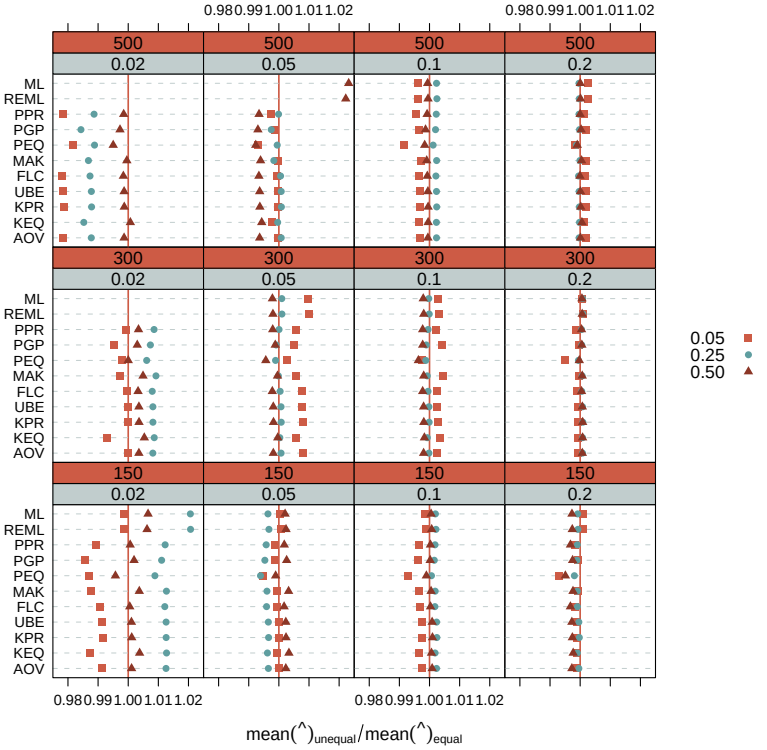


Figure 49: *Grouped dotplots of means of estimators of ρ under two-stage cluster sampling with unequal to equal cluster sizes for binary data*

on variable cluster sizes compared to the scale of equal cluster sizes – especially so when cluster sizes are small.

Figure 50 follows the same logic but now displays the ratios of standard deviations of estimators under unequal to equal cluster sizes instead of means. What can be seen at first glance is that many dots lay above one, indicating that many estimators in many settings react with a loss in precision on a switch from equal to unequal cluster sizes.

The $\hat{\rho}^{(\text{KEQ})}$, $\hat{\rho}^{(\text{MAK})}$, and $\hat{\rho}^{(\text{PGP})}$ estimators lose precision when cluster sizes vary compared to settings where they are constant. This effect is, again, amplified by small population parameter of ρ and small cluster sizes. Especially $\hat{\rho}^{(\text{KEQ})}$ shows less loss in precision when the parameter to estimate is large. $\hat{\rho}^{(\text{PEQ})}$, however, behaves different: with an increase in population ρ a switch from equal to unequal cluster sizes increases the variation of this estimator. In addition, a decrease in cluster size decreases the influence of this effect – at a given level of ρ, smaller cluster sizes have a positive effect on the relative precision of $\hat{\rho}^{(\text{PEQ})}$. The estimators of ρ which are based on variance components of a random effects model react sensitively to a

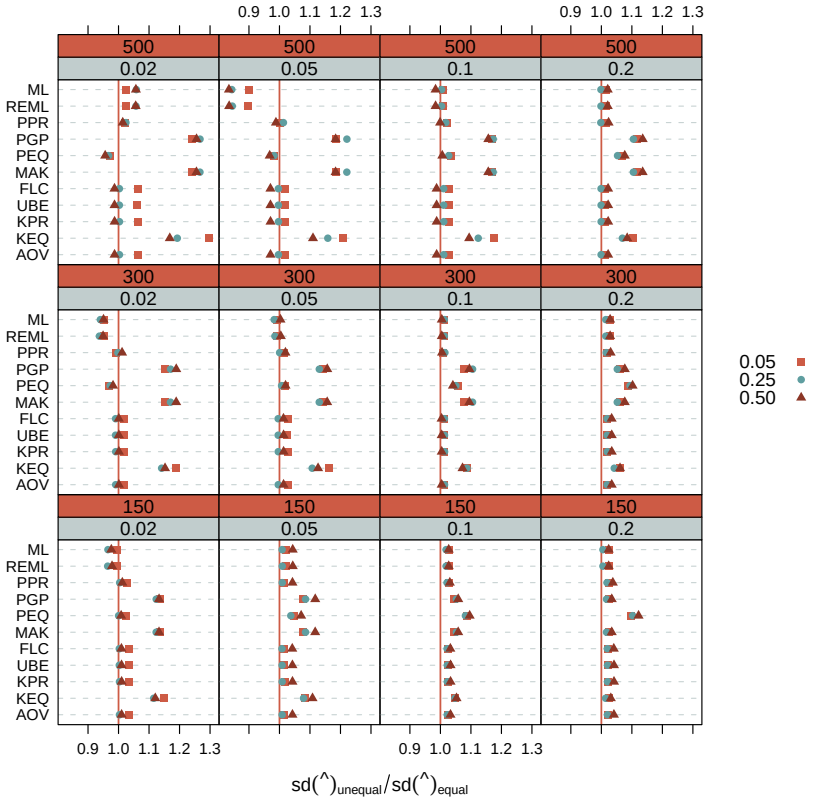


Figure 50: *Grouped dotplots of standard deviations of estimators of ρ under two-stage cluster sampling with unequal to equal cluster sizes for binary data*

switch from equal to unequal cluster sizes only in extreme scenarios, namely when population ρ and cluster sizes are small. Then, however, both $\hat{\rho}^{(\mathrm{ML})}$ and $\hat{\rho}^{(\mathrm{REML})}$ gain from such a shift – these estimators are more precise with varying than with constant cluster sizes.

5.6 Comparison of Estimation Strategies

With a given complex sample at hand, the most interesting question to answer for the data analyst is: *Which estimator of the design effect fits best my data?* Depending on the properties of the data and the study variable, the answer to this question may vary. This section gives a comparative round-up of estimation approaches for different settings, scenarios, approaches and estimators. The comparison is based on three basic summary statistics: the mean, standard deviation (sd) and coefficient of variation (cv) of estimated design effects over iterations. Each evaluation of the quality of point estimation uses the Monte Carlo estimated true design effect as a reference. The visualization is by use of grouped dotplots. Within a panel, summary

statistics are displayed for estimators of each approach (design-based and model-based approach). Estimators are distinguished by different plot symbols. Estimates of model-based estimators are calculated according to formula (3.5), substituting ρ by one of its estimators. It can be seen that all estimators yield very similar mean estimates.

5.6.1 Continuous Data

This subsection summarizes and compares the results of three previous sections on a) the Monte Carlo estimated true design effect, b) design-based and c) model-based estimation of the design effect for continuous data. Due to its bias, the $\hat{\rho}^{(\text{Laplace})}$ estimator is not considered any further in the following discussion.

5.6.1.1 Cluster Sampling with equal Cluster Sizes

Starting with the evaluation of differences in point estimation, Figure 51 gives an overview of the distribution of mean estimated design effects by approach and estimator for different scenarios. There is hardly any difference between estimation

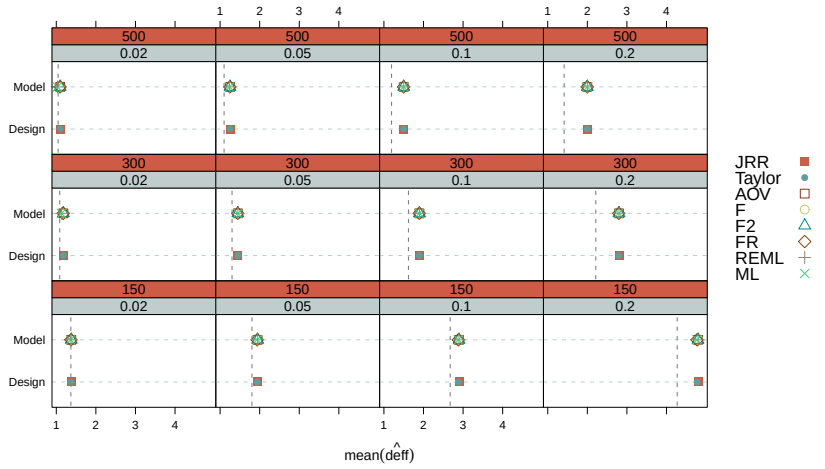


Figure 51: *Grouped dotplots of the mean of $\widehat{deff}$ under cluster sampling with equal cluster sizes for given scenarios with continuous data*

approaches. Compared with the Monte Carlo estimated true design effect, all estimators (except for the one based on $\hat{\rho}^{(\text{Laplace})}$) yield conservative estimates (i.e. the mean estimated design effect is bigger than the Monte Carlo estimated true design effect in the same scenario) which is indicated by the location of the plotting characters right to the vertical dashed line.

Let us take a look at the precision of the estimators. Figure 52 shows the standard deviations of estimators over 10 000 iterations of a scenario. Estimators of either approach are very similar in terms of precision. The model-based estimators using $\hat{\rho}^{(\text{ML})}$ and $\hat{\rho}^{(\text{REML})}$ as estimators for ρ are, however, less precise than all other estimators –

especially when the population parameter to estimate and the cluster sizes are rather small.

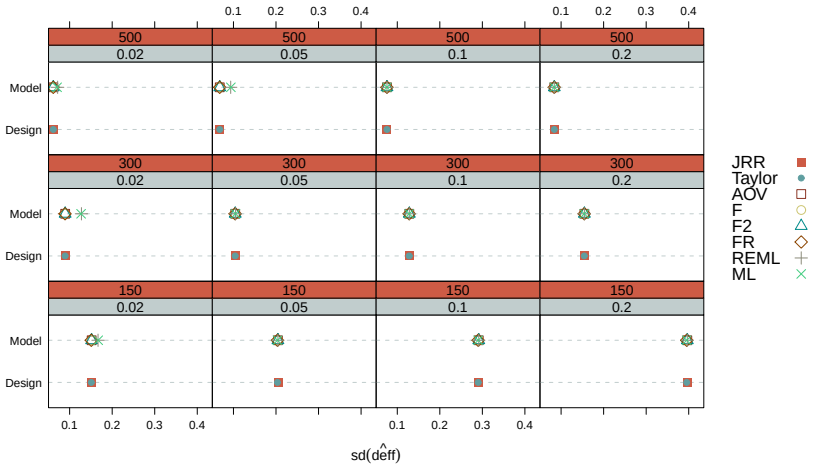


Figure 52: *Grouped dotplots of the standard deviations of $\widehat{deff}$ under cluster sampling with equal cluster sizes for given scenarios with continuous data*

5.6.1.2 Cluster Sampling with unequal Cluster Sizes

If variation in cluster sizes is present and hence weighting comes into play, estimators of *deff* do have to include the additional inflation in variance due to weighting. The design-based estimators do so by directly incorporating weights in the variance estimator while model-based estimators are ex-post *calibrated* by multiplying $\widehat{deff}_c$ with $\widehat{deff}_p$. The point estimates of all estimators behave very similar also when weighting is present and will thus not be reported here. However, estimators of *deff* differ in terms of precision as Figure 53 shows. Again, model-based estimators of *deff* applying $\hat{\rho}^{(\mathrm{ML})}$ and $\hat{\rho}^{(\mathrm{REML})}$ as estimators for ρ tend to be less precise. What is more interesting, however, are the differences between estimators of different approaches. The design-based estimators ($\widehat{deff}^{JRR}$ and) of *deff* are less precise than the most precise model-based estimator based on $\hat{\rho}^{(\mathrm{FR})}$ in all settings. In fact the JRR estimator is 2.3% ($\rho = 0.20$; $m = 150$) to 24.6% ($\rho = 0.02$; $m = 500$) less precise than the AOV model-based estimator. The design-based $\widehat{deff}^{Taylor}$ estimator behaves very similarly: its precision is 1.7% to 24.4% less that that of the model-based AOV estimator of *deff*.

5.6.2 Binary Data

Turning to binary data, a further parameter has to be considered in the comparisons, namely the overall rate of success of the study variable, π. In order to keep interpretation of the following plots as straightforward as possible, each of them compares the estimators bias and precision for given levels of m.

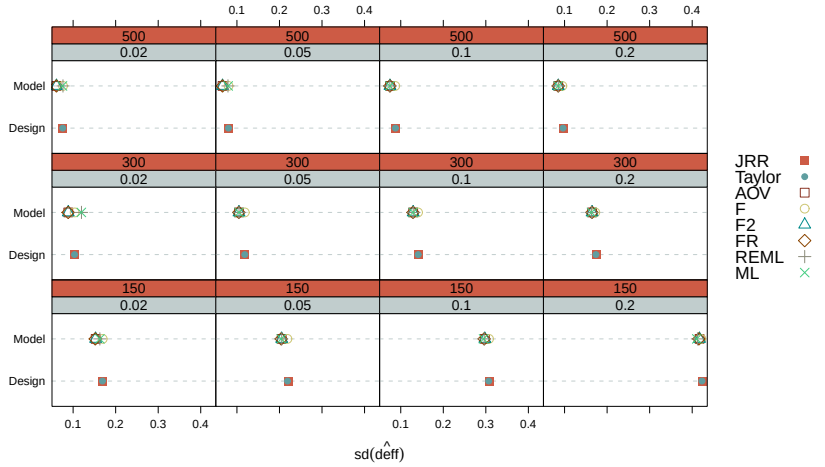
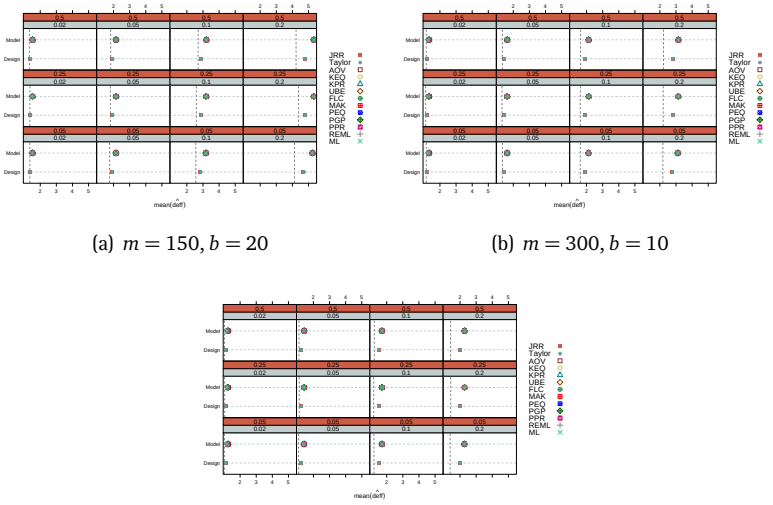


Figure 53: *Grouped dotplots of the standard deviations of $\widehat{deff}$ under cluster sampling with unequal cluster sizes for given scenarios with continuous data*

5.6.2.1 Cluster Sampling with equal Cluster Sizes

In terms of bias, the differences in the means of estimators between approaches are more obvious than in the previous setting as Figure 54 demonstrates.



(a) $m = 150, b = 20$ (b) $m = 300, b = 10$

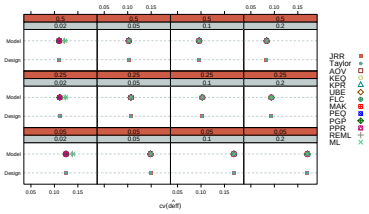
(c) $m = 500, b = 6$

Figure 54: *Grouped dotplots means of estimates of $\widehat{deff}$ under cluster sampling with equal cluster sizes for given scenarios with binary data*

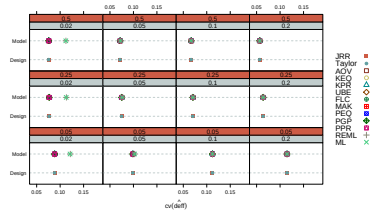
What can be seen at first glance is that the design-based estimators are on average much closer to the Monte Carlo estimated true design effect than the model-based

estimators. With equal cluster sizes, however, there seems to be hardly any effect of π on the magnitude of this difference but rather an effect of the population parameter of ρ: in highly clustered universes, model-based estimation techniques are over-conservative in respect to the Monte Carlo estimated design effect.

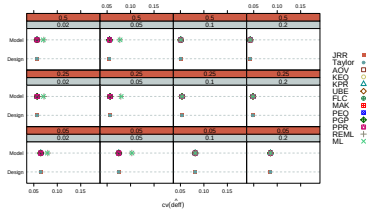
Due to the influence of b on the magnitude of $\widehat{deff}$, also the variance of the estimates will be influenced by changes in b. This is why Figure 55 shows the coefficients of variation of estimators in different scenarios.



(a) $m = 150, b = 20$



(b) $m = 300, b = 10$



(c) $m = 500, b = 6$

Figure 55: *Grouped dotplots of coefficients of variation of estimates of $\widehat{deff}$ under cluster sampling with equal cluster sizes for given scenarios with binary data*

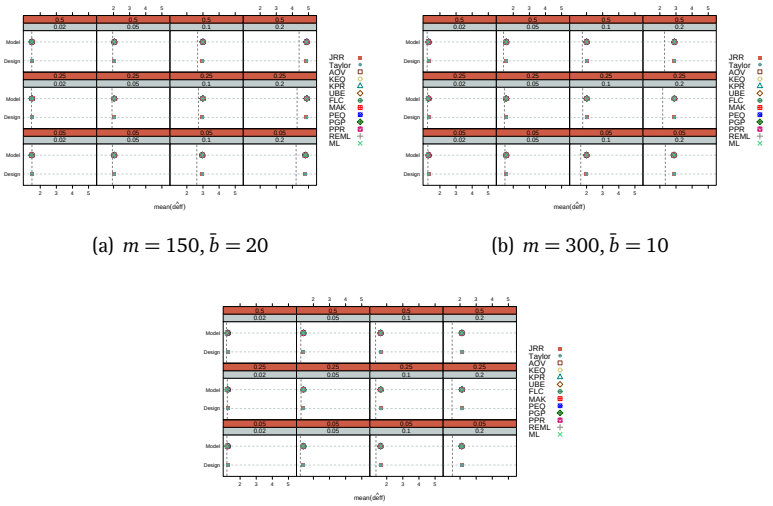
This figure illustrates that all estimators are less precise when π is small and when ρ in the population is large as can be seen from a comparison of the dots between panels in a column and in a row. This is true for all levels of m. A decrease in the cluster size, on the other hand, has a positive effect on the precision. This indicates that the dominating factor ruling the precision of any estimator is the skewness of the study variable and the degree of homogeneity in the population.

5.6.2.2 Cluster Sampling with unequal Cluster Sizes

Also with binary data, variation in the size of sampled clusters makes weighting necessary. This will increase the variance of the HT estimator and hence the design effect in the same way described previously.

Figure 56 shows that with unequal cluster sizes and in the presence of weighting the differences between model-based and design-based estimators diminish. Within one panel of the plot, the means of all estimators lay very close to each other – there are hardly any differences in magnitude between the two approaches. The conserva-

tive nature of the estimators compared to the Monte Carlo estimated true design effect (dashed vertical line), however, is also present with unequal cluster sizes.



(a) $m = 150, \bar{b} = 20$ (b) $m = 300, \bar{b} = 10$

(c) $m = 500, \bar{b} = 6$

Figure 56: *Grouped dotplots of means of estimates of $\widehat{deff}$ under cluster sampling with unequal cluster sizes for given scenarios with binary data*

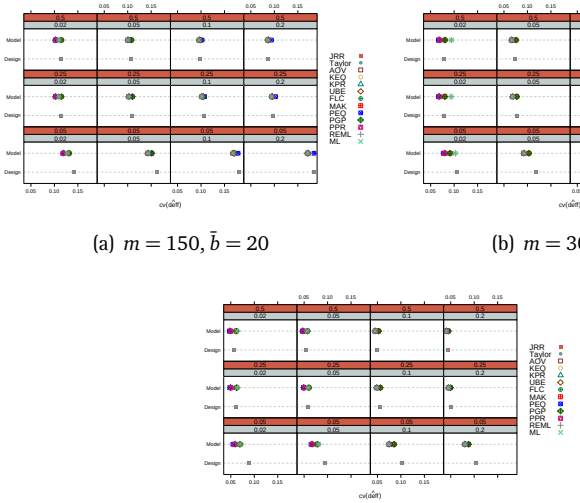
In line with what could be observed in the previous subsection, there is hardly any effect of π on the point estimates also when cluster sizes vary and weighting comes into play.

Things, however, change when we look at the estimators' precision. As before, the coefficient of variation serves as measure in Figure 57. An overall pattern which can be observed is a decrease of the coefficient of variation as π gets larger (i.e. the study variable is less skewed) and as the population parameter of ρ gets smaller. A decrease in average cluster size also has a positive effect on precision of the estimators. The design-based estimators tend to be less precise than most of the model-based estimators[20] when π is small. With $\pi = \{0.50, 0.25\}$ all estimators tend to be more precise as the population parameter of ρ increases (comparison of plot symbols in a row).

5.7 Decomposition of Design and Interviewer Effects

In face-to-face sample surveys with a cluster sample design, one is faced with the problem that variance on the response variable can be attributed to three sources: 1.) the geographical cluster, 2.) the interviewer and 3.) residual variance. In such a situation, estimation of ρ_{PSU} will be influenced by the additional source of homogeneity

20 Except for the estimator based on $\hat{\rho}^{(\text{PEQ})}$ in the scenarios ($m = 150, \rho = 0.2$), ($m = 300, \rho = 0.2, \pi = \{0.25, 0.50\}$) and the estimators based on $\hat{\rho}^{(\text{ML})}$ and $\hat{\rho}^{(\text{REML})}$ in the scenarios ($m = \{300, 500\}, \rho = 0.02, \pi = \{0.25, 0.50\}$)



(a) $m = 150, \bar{b} = 20$ (b) $m = 300, \bar{b} = 10$

(c) $m = 500, \bar{b} = 6$

Figure 57: *Grouped dotplots of coefficients of variation of estimates of $\widehat{deff}$ under cluster sampling with unequal cluster sizes for given scenarios with binary data*

introduced by the interviewer. When ignoring the additional level of clustering, $\hat{\rho}^{(\bullet)}_{\text{PSU}}$ tend to be biased. The magnitude and direction of bias depends on a) the structure of nesting or crossing of interviewer with geographical clusters, b) on the magnitude of ρ_{PSU} and ρ_{INT} as well as on c) the share of each variance component on the total variance.

For this simulation study a simple nested structure of two interviewers in one geographical cluster was assumed for simplicity (see Section 5.1.2 for details). For reasons of computation time, only 2 500 iterations per parameter combination for continuous data have been performed in this simulation study. As before, results refer to the HT estimator of the population mean. With the nested structure, one has to distinguish additionally between m_{PSU} and m_{INT}. As the generation of the universes is such that interviewer clusters within PSUs are perfectly balanced, $m_{\text{INT}} = 2 \times m_{\text{PSU}}$.

5.7.1 Estimation of Intraclass Correlation with Nested Data

This subsection presents the effects of naive estimation of $\hat{\rho}^{(\bullet)}_{\text{PSU}}$ in given scenarios. Naive means that the estimation of intraclass correlation is on the PSU level although additional clustering at an interviewer level is present. Due to the additional dimension introduced by the two parameters of ρ_{INT} and s_{INT}, the following figures show results for levels of s_{INT} (rows) and ρ_{INT} at given levels of m_{PSU} (and associated with that $m_{\text{INT}} = 2 \times m_{\text{PSU}}$) and ρ_{PSU}.

5.7.1.1 Equal cluster sizes

The following figures show the distribution of $\hat{\rho}_{\text{PSU}}^{(\bullet)}$ based on repeated draws from populations with ρ_{PSU}=0.02 (Figure 58) and ρ_{PSU}=0.10 (Figure 59) with $m_{\text{PSU}} = 150$; $m_{\text{INT}} = 300$ each for levels of ρ_{INT} and s_{INT}. Both figures show the influence of changes in s_{INT} (comparison of boxplots in a row) and in ρ_{INT} (comparison of boxplots in a column).

With ρ_{PSU}=0.02 (Figure 58) before mixture (see equation (5.2) in Section 5.1.2 on page 60), we see at first glance that the differences between estimators are rather small. Turning to the effects of s_{INT} and ρ_{INT}, we can observe a clear pattern: with

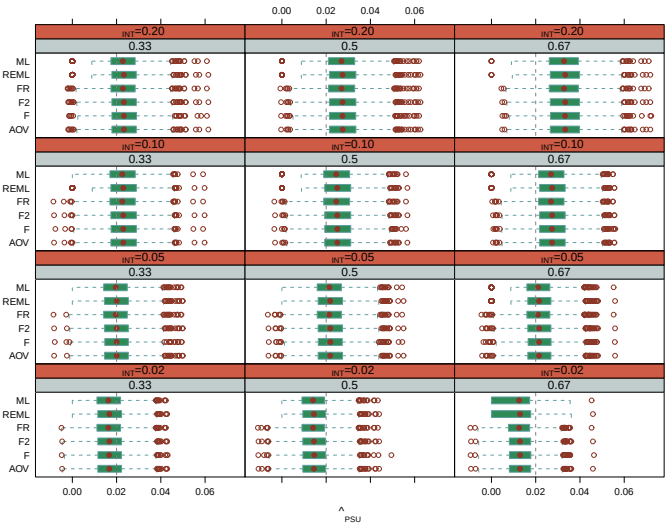


Figure 58: *Estimated ρ_{PSU} by levels of ρ_{INT} and s_{INT} for $m_{PSU} = 150, m_{INT} = 300$ and population $\rho_{PSU} = 0.02$ for equal cluster sizes*

s_{INT}=0.33 (i.e. 33% of the total variance are attributed to the interviwer clusters), the effect of an increase in ρ_{INT} is rather small. In fact, most estimators of ρ_{PSU} are downwards biased when ρ_{INT} =0.02; upward-bias is present with $\rho_{\text{INT}} = \{0.10, 0.20\}$. With an equal share of PSU and interviewer clusters on the variance components (s_{INT}=0.50), the downward-bias (at small levels of ρ_{PSU}) and the upward-bias (at all other levels of ρ_{INT}) is more obvious. This tendency is even more pronounced at s_{INT}=0.67.

Similar patterns can be observed in Figure 59, which shows results for a setting where ρ_{PSU} in the population before mixture is 0.10. Here the effects of both, s_{INT} and ρ_{INT}, are even stronger. All estimates of all estimators in the panels of the lower row lay below the initial population value of ρ_{PSU} indicating a strong effect of s_{INT} and of ρ_{INT} during the process of mixing. Compared to the previous setting, the influence of ρ_{INT} at any given level of s_{INT} is even stronger, leading to greater changes

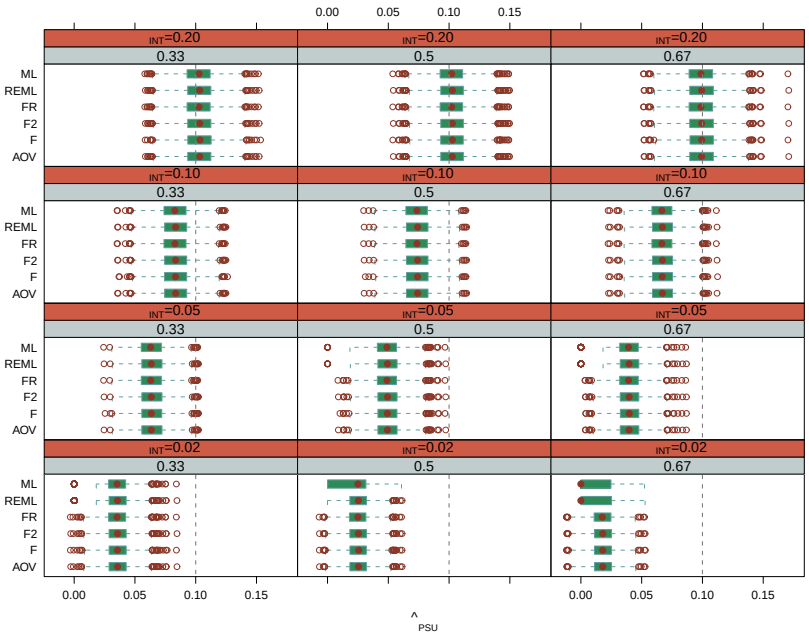


Figure 59: *Estimated ρ_{PSU} by levels of ρ_{INT} and s_{INT} for $m_{PSU} = 150, m_{PSU} = 300$ and population $\rho_{PSU} = 0.10$ for equal cluster sizes*

between the boxplots of a column. The effect of an increase in s_{INT} at a given level of ρ_{INT}, however, is weaker than before. Hence, bias in the estimation of the initial population parameter is positively correlated with ρ_{PSU}. A researcher who naively estimates ρ_{PSU} ignoring homogeneity introducing by interviewers will generally get to biased results. The direction and the degree of bias will depend on the magnitude of interviewer homogeneity, ρ_{INT}, and the share that it has on the overall explained variance, s_{INT}.

5.7.1.2 Unequal Cluster Sizes

The following figure shows the distribution of estimators of ρ_{PSU} based on repeated draws from populations with $\rho_{\text{PSU}} = 0.02$ (figure 60) for levels of ρ_{INT} and s_{INT}. Due to the fact that the patterns in the distributions of the estimators are very similar to the scenarios presented in the previous subsection, this setting only reports results for the setting ($\rho_{\text{PSU}} = 0.02, m_{\text{PSU}} = 150, m_{\text{INT}} = 300$).

The distribution of the expected values of estimators and the patters of influence of the parameters which are varied are very similar to the setting with equal cluster sizes. With unequal cluster sizes, however, differences in the estimators' precision (i.e. spread of the boxes and outliers marked by red hollow dots) become more visible.

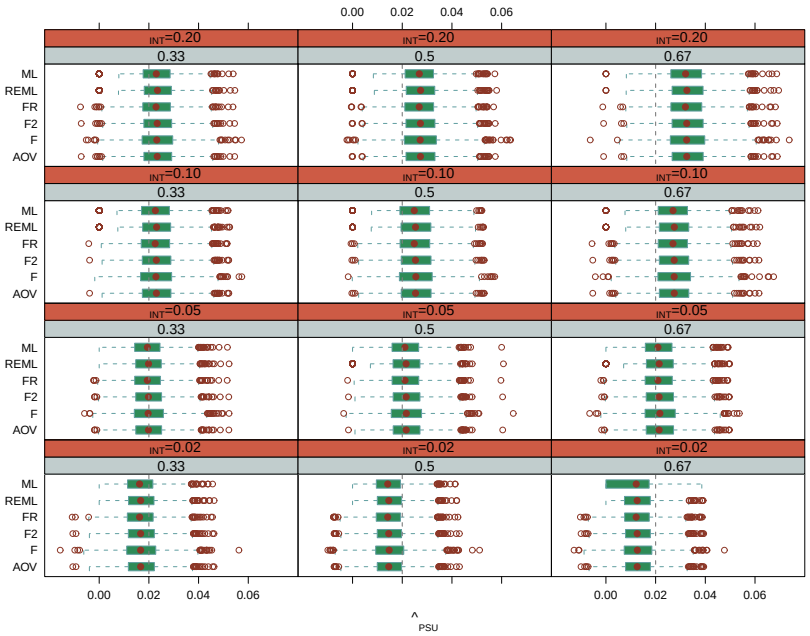


Figure 60: *Estimated ρ_{PSU} by levels of ρ_{INT} and s_{INT} for $m_{PSU} = 150, m_{INT} = 300$ and population $\rho_{PSU} = 0.02$ for unequal cluster sizes*

5.7.2 Variance Decomposition

For the variance decomposition, first a random effects model was fitted on the data with PSU and INT as nested random effects. Then, a variance decomposition was performed using the VarCorr() function of the R package lme4. Finally, the ratio of the variance component to the interviewer level to the total variance of the model was calculated as $s_{\text{INT}} = \dfrac{\hat{\sigma}_{\text{INT}}}{\hat{\sigma}_{\text{INT}} + \hat{\sigma}_{\text{PSU}}}$. This procedure was executed for each of the 2 500 iterations in each scenario.

5.7.2.1 Equal Cluster Sizes

With equal cluster sizes the patterns of the estimation of the share of the variance due to the interviewer are relatively stable as Figure 61 indicates. It shows mean estimated s_{INT} in different scenarios. The main factor of influence within any given panel is the cluster size. Estimation of the model variance components using the VarCorr() function of R reacts sensitive to cluster sizes with rather upwards biased results when average cluster sizes are small and rather downwards biased mean estimates when cluster sizes are large. The magnitude of both ρ_{PSU} and ρ_{INT} in the population of course also influence the point estimates. An increase of ρ_{PSU} (vertical comparison within a panel) generally leads to less bias – this is especially true for small cluster sizes and medium and large values of s_{INT}. When ρ_{INT} is small and the share of the

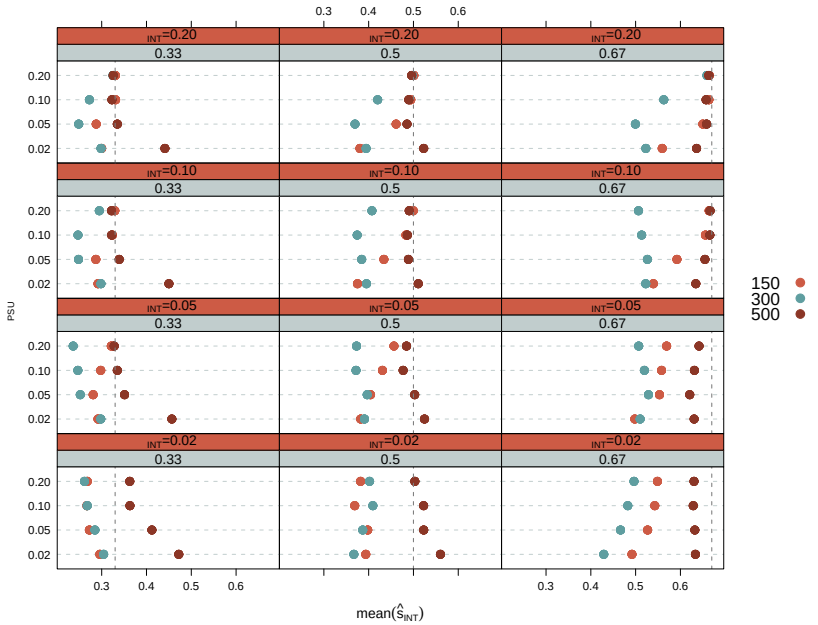


Figure 61: *Dotplot of estimated mean s_{INT} by levels of ρ_{INT} and m with equal cluster sizes*

interviewer component is 0.67 (lower right panel) downwards bias is most severe for all levels of cluster sizes.

When it comes to precision we can observe different patterns. Figure 62 shows the distribution of standard deviations of estimates in different scenarios. An overall pattern is that precision decreases as both ρ_{PSU} and ρ_{INT} decrease (comparison within a row, within a column and in combination). The standard deviation also increases with an increase of the share of interviewer variance in the population. This effect, however, could also be influenced by the magnitude of the parameter to estimate. However, the pattern is still present if we standardize on the mean (i.e. regard the coefficient of variation). The VarCorr() function also reacts sensitive to small cluster sizes, especially when overall interviewer homogeneity is small (lower row). Patterns are less clear cut for larger values of ρ_{INT}, however.

5.7.2.2 Unequal Cluster Sizes

Turning to unequal cluster sizes we must, as in other situations, consider design weights also in the estimation of the random effects model and the extraction of variance components. As could be expected, design weighting has hardly any effect on point estimates as can be seen from Figure 63. The distribution of means of estimated s_{INT} shows very similar patterns as in the case of equal cluster sizes and hence without weighting.

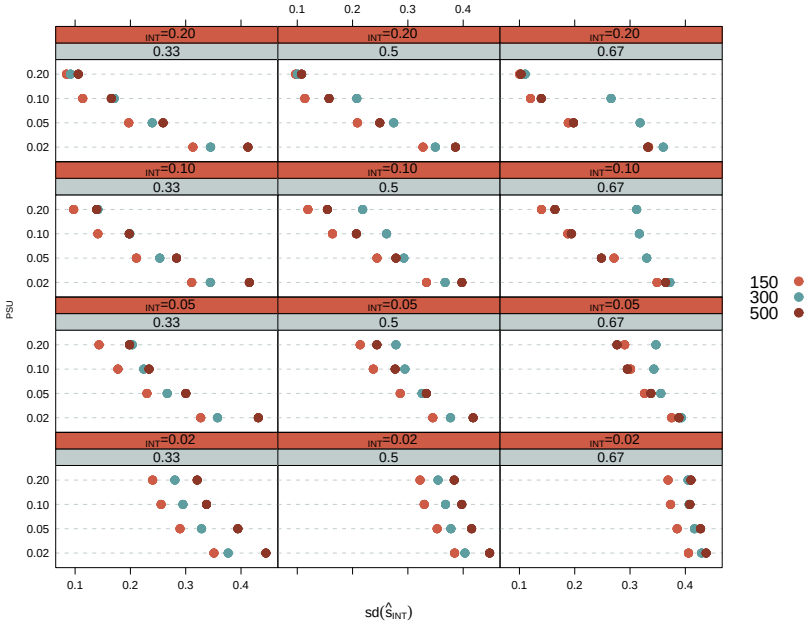


Figure 62: *Dotplot of standard deviation of estimated s_{INT} by levels of ρ_{INT} and m with equal cluster sizes*

A look at Figure 64 also reveals no other patterns than in the case of equal cluster sizes.

5.7.2.3 Comparison of Cluster Sampling with equal and unequal Cluster Sizes

A comparison of the point estimates of s_{INT} under cluster sampling with unequal and with equal cluster sizes is given in Figure 65 which shows the ratios of mean estimated s_{INT} under cluster sampling with unequal to equal cluster sizes. A pattern that can be observed is the decrease in variation of the ratios of point estimates as ρ_{PSU} and ρ_{INT} increase. Average cluster size seem to have a clear cut effect only when ρ_{INT} is small and s_{INT} is medium or large.

The ratios of the standard deviations of estimates under cluster sampling with unequal and equal cluster sizes, shown in Figure 66, are a bit of a surprise, however. One would expect the precision of estimated variance components to be lower if additional variance is introduced in the estimation by design weights. The ratios of standard deviations shown in the above figure, however, indicate that in some scenarios design weighting even increases the precision (i.e. some dots are left to the dashed vertical line, indicating that the standard deviation in the corresponding scenario is smaller than the one without design weighting). Strangely this effect is stringer when average cluster sizes are small.

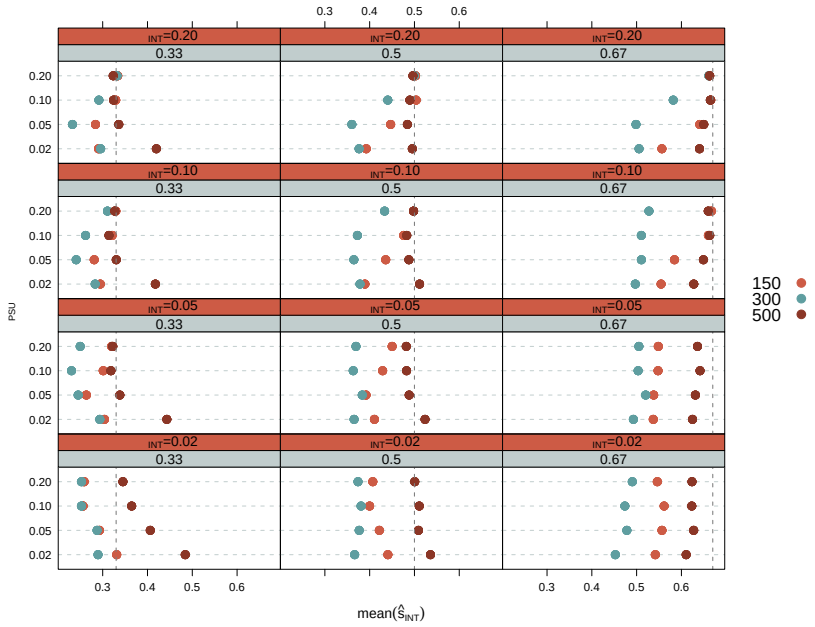


Figure 63: *Dotplot of estimated mean s_{INT} by levels of ρ_{INT} and m with unequal cluster sizes*

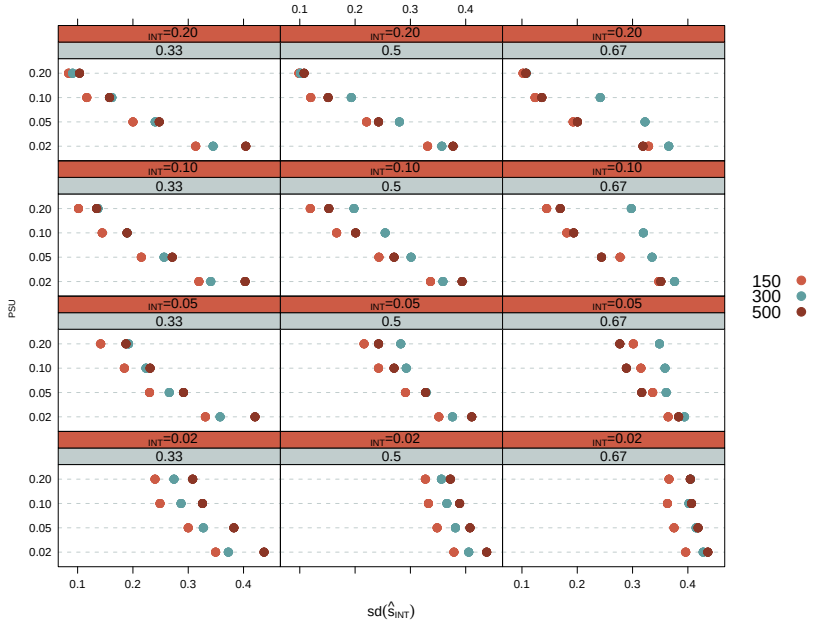


Figure 64: *Dotplot of standard deviation of estimated s_{INT} by levels of ρ_{INT} and m with unequal cluster sizes*

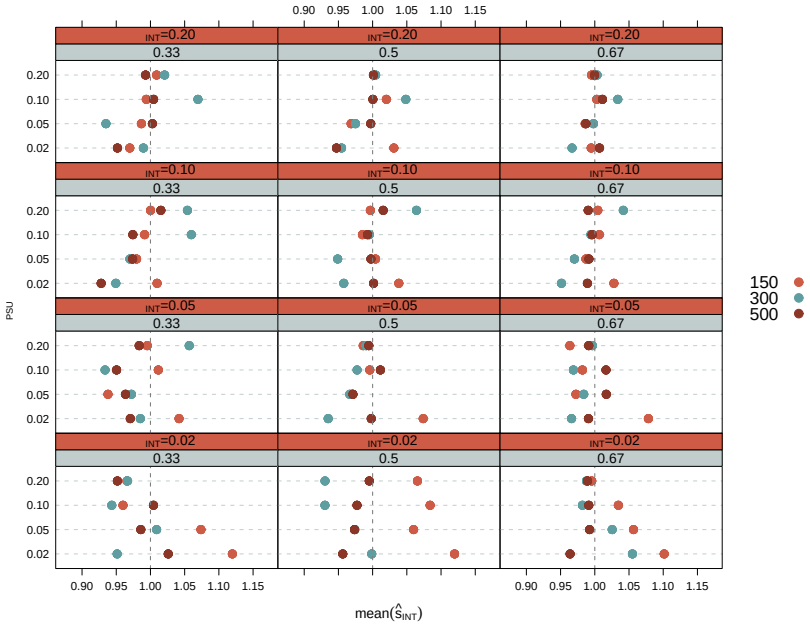


Figure 65: *Dotplot of ratios of estimated mean s_{INT} by levels of ρ_{INT} and m with unequal to unequal cluster sizes*

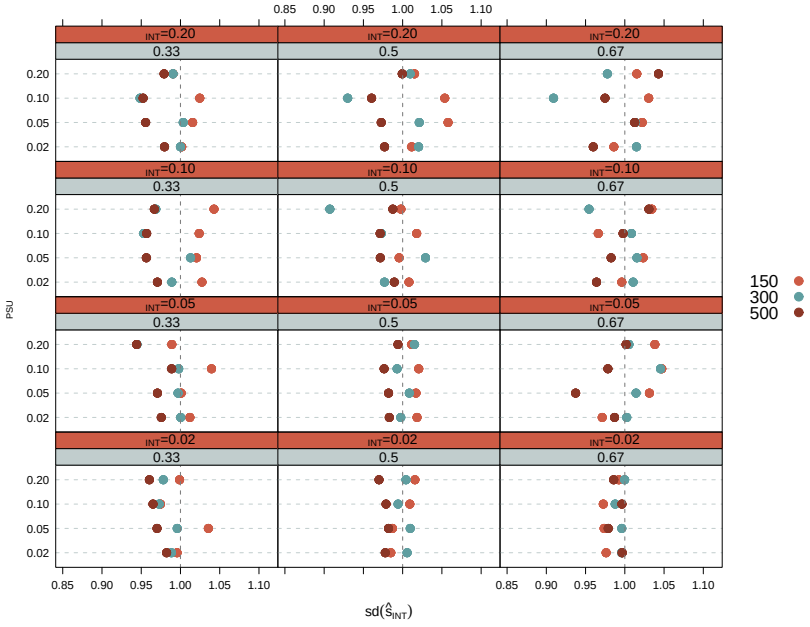


Figure 66: *Dotplot of ratios of standard deviations of s_{INT} by levels of ρ_{INT} and m with unequal to unequal cluster sizes*

6 Estimation of Design Effects in the European Social Survey

This chapter links the findings of the simulation studies presented in the previous chapter with a real-world complex sample survey. The European Social Survey (ESS) is known for its methodological and statistical rigour. It explicitly employs the concept of design effects for planning purposes. In light of restricted budgets, however, the development of ESS sampling strategies for participating countries must consider both, quality and cost issues. Thus, estimation and prediction of design effects and their components must be able to rely on highly precise estimators.

In the following Section 6.1 gives an overview of the ESS project as a whole and of the work of the sampling expert panel in specific. The general procedure which relies on the model-based approach to plan the sample designs of participating countries is described in section 6.2. Then, in Section 6.3 sample designs of selected countries are presented and their implications in terms of design effects are discussed. Finally, Section 6.4 presents the estimation of design effects and illustrates how strategic planning of sample designs has an influence on their magnitude.

6.1 Aim and overall Design of the ESS

"The European Social Survey (the ESS) is an academically-driven social survey designed to chart and explain the interaction between Europe's changing institutions and the attitudes, beliefs and behaviour patterns of its diverse populations. Now in its fourth round, the survey covers over 30 nations and employs the most rigorous methodologies. The survey has been funded through the European Commission's Framework Programmes, the European Science Foundation and national funding bodies in each country."

http://www.europeansocialsurvey.org

The management of the ESS is done by a so called Central Coordinating Team (CCT), an international group of experts in various fields of methodological research. Members of the CCT contribute their knowledge to workpackages which aim at supporting central routine tasks the ESS is faced with each round. The so called sampling expert panel is a group of five experts in the field of sampling who, in cooperation with the National Coordinator, develop a sample design to be applied in a given country and round. This development is guided by a principle that meanwhile became widely accepted in cross-national sample survey research. It states that "workable and equivalent sampling strategies in all participating countries" have to be developed (Häder et al., 2007, 2). This goal is to be seen in line with Kish (1994, 173):

"Sample designs may be chosen flexibly and there is no need for similarity of sample designs. Flexibility of choice is particularly advisable for multinational comparisons, because the sampling resources differ greatly between countries. All this flexibility assumes probability selection methods: known probabilities of selection for all population elements."

Nevertheless, there are some minimum requirements that any sample design has to meet to find unanimous acceptance among the members of the sampling expert panel.

6.2 Using the Model-based Approach to predict required sample sizes

In the ESS, the sampling team uses a model-based approach to predict the required number of achieved interviews (i.e. n_{net}) to reach a pre-defined effective sample size of $n_{eff} = 1\,500$. Ideally, if this number was achieved in all participating countries, differences in the variance of an estimator would be independent of the sample design. However, the effective sample sizes vary between countries and variables under study as both, the net sample size (due to item non-response) and also design effects (also due to item non-response but also due to the magnitude of $\hat{\rho}$) are subject to variation even within a given country. Also between countries, however, there is variation in n_{eff} within the same study variable as n_{net} and n_{gross} are also heavily depending on the budget available.

6.3 Sample Designs in selected Countries

A set of four countries is considered in more detail in the following. This helps limiting the complexity of comparison while at the same time maximizing the variation in sample designs. Hence, the countries under study first apply most various sample designs but should further have taken part in at least two of the three ESS rounds that have been conducted so far in order to spot changes in the sample quality if meaningful alterations have been applied to the sample design. For this reason, I choose Spain (ES), Finland (FI), France (FR), and Poland (PL) for the following analysis.

In this selection, Spain is a typical candidate for a country with a stratified two-stage sample design. In such a design, usually at the first stage municipalities (primary sampling units) are selected with probability proportional to size. Then, at the second stage, a fixed number of persons is sampled from a complete list (e.g. administrative records, electoral registers, residential population register, etc.) by srs.

If not persons but only super-ordinate sampling elements (e.g. households or addresses) are elements of the list available at the second stage, these secondary sampling units (SSU) make up an additional stage of selection. In this scenario, after the selection of a fixed number of households has taken place, in each SSU one respondent is selected using appropriate techniques (e.g. last/next birthday or a Kish grid).

France, in turn is a representative of the set of countries with an area-based selection of households. Here, random-route procedures are applied to generate a list of households to contact. In a different version of this procedure a complete list of households or addresses is compiled by a fieldwork person. Then, this list is returned to the fieldwork agency and a simple random sample of households is drawn.

In Poland, the sample is drawn independently in two separate domains. Usually, one domain consists of big cities where lists of ultimate sampling units are available. In this domain, a simple random or a stratified sample of ultimate sampling units is

drawn directly. In the second domain, which usually covers settlements that do not belong to the first domain (i.e. villages and small towns in rural areas), a two- or three-stage probability sample is drawn.

Finland is one of the countries in the ESS where single-stage simple random or stratified random sampling can be applied directly. Hence, no clustering is involved in achieving the sample.

The basic structure of sample designs in the ESS is captured on a micro level in so called *sample design data files* (SDDF). These SDDF include several important variables which characterize the sample design, for example first order inclusion probabilities at each stage (variables PROB1 to PROB4) or the PSU label an individual belongs to. These files are generated accompanying the fieldwork process and are delivered to the sampling expert panel for further process (e.g. generating design weights or estimation of design effects as a basis for the upcoming round). The SDDF are not generally publicly available. However, privacy regulations in some countries allow publication as long as anonymity of respondents can be assured.

In the following subsections, I describe the sample designs of the selected countries of round 1 to 3 of the ESS in more detail. I will emphasize on the effects of deliberate changes and describe the effects of flaws and variations from the prescribed sample designs.

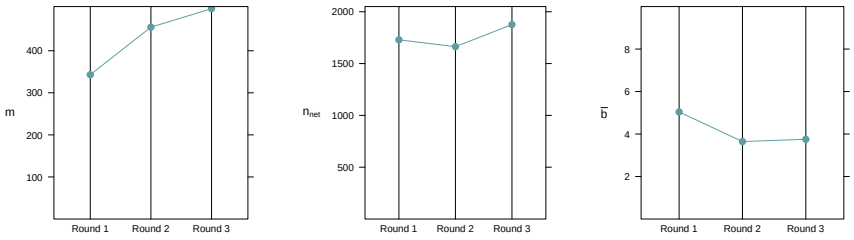
6.3.1 Spain

The whole population of Spain is divided into 33 000 electoral sections[21] which are provided as a frame through the so called *municipal roll*. The complete frame, however, is not available for scientific research but only a *master sample*, a sub-sample of 3 500 electoral sections which is updated continuously.

From the master sample, m electoral sections are allocated to H strata. The Cox-Method of controlled rounding is used to ensure a fixed sample size. Then, at the first stage $h = 1,\ldots,H$, m_h electoral sections (PSUs) are sampled with probability proportional to population size in each stratum[22]. At the second stage, a fixed number of individuals is selected via simple random sampling. In round 1, however, there was not access to a list of individuals at the PSU level but only to lists of households. Hence, at the second stage, instead of individuals, households were selected and the selection of persons was shifted to the third stage and was done via the last birthday method. The dotplots of figure 67 illustrate the development of m, n_{net} and $\bar{b}$ over the rounds. As can be seen, Spain has increased the number of PSUs from round 1 to round 2 and 3 from $m_{R1} = 343$ to $m_{R2} = 456$ and $m_{R3} = 500$ which resulted in an overall decline in the average cluster size from $\bar{b}_{R1} = 5.04$ to $\bar{b}_{R2} = 3.65$ to $\bar{b}_{R3} = 3.75$. This decline can be expected to have an influence on the magnitude of $\widehat{deff}_c$ as $\hat{\rho}$ will not be very likely to go down since the composition of elements in PSUs remains

21 As from ESS round 1 (2002). In round 2 (2004), the municipal roll was updated and contained 34 600 electoral sections.

22 Note that in round 1, proportionality was with respect to households, not persons. In round 2 and 3, inclusion probabilities were based on the size of the target population.



(a) m (b) n_{net} (c) $\bar{b}$

Figure 67: *Development of m, n_{net} and $\bar{b}$ in Spain over round 1, 2 and 3*

the same. This development is rather typical for a country with a multi-stage sample design. As there is no alternative frame available at the moment, the priority is to design the sample in such a way that *deff* will not be too large. This can be achieved most easily with an increase in the number of sampled clusters and hence a decrease in $\bar{b}$.

For round 1, where a three-stage design was applied, the product of inclusion probabilities of the first two stages is constant[23]. Hence, the only variation in first order inclusion probabilities, π_i=PROB1×PROB2×PROB3, can be attributed to variation at the third stage. Here, inclusion probabilities directly depend on the number of individuals in a household who belong to the target population. Thus, also (normalized) design weights will vary only to the degree to which household sizes vary. As an effect, also $\widehat{deff}_p$, which only depends on the distribution of weights and n_{net}, will be of reasonable magnitude[24]. The distribution of normalized design weights in round 1, 2 and 3 is depicted in the grouped densityplot of figure 68. Here, one can easily see that normalized design weights in round 1 are much wider spread than in round 2 and 3. This is due to the fact that variation in inclusion probabilities, and hence in design weights, in round 2 and 3 is only due to deviations from the planned number of sampled individuals, for example due to non-response. Ideally, without any distortion, the product of inclusion probabilities would be equal for the sample design applied in these rounds.

6.3.2 France

Unlike in Spain, where access to population registers is granted to scientific bodies, privacy regulations in France do not permit access to such registers (although they do exist). This is why in all three rounds conducted so far, a stratified multi-stage

23 This is due to the fact that for the ith PSU, PROB1 $= \frac{N_i}{N}$ and PROB2 $= \frac{c}{N_i}$ and PROB1 × PROB2 $= \frac{N_i}{N} \times \frac{c}{N_i} = \frac{c}{N}$.

24 Note that with a multi-stage sample where the only source of variation in inclusion probabilities, and hence in design weights, stems from the variation in household sizes, the magnitude of $\widehat{deff}_p$ is typically 1.2 to 1.3

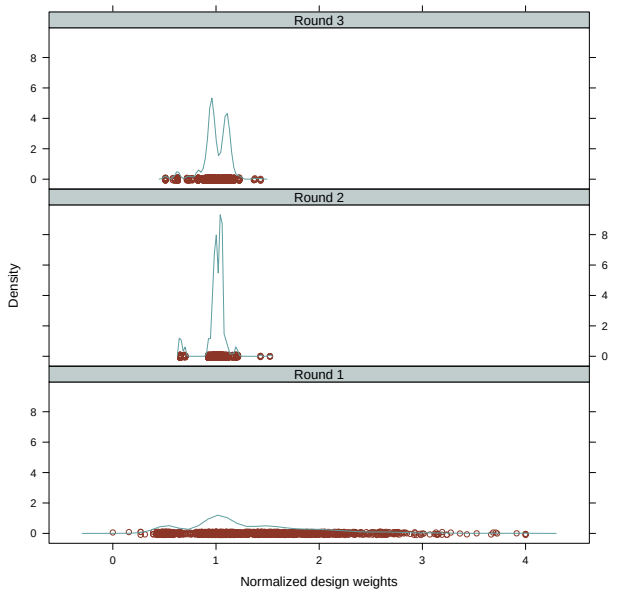


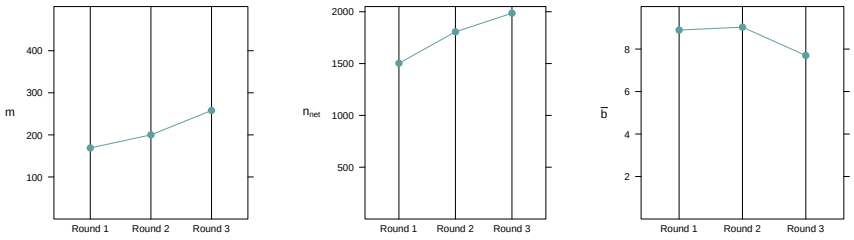
Figure 68: *Grouped density plot of normalized weights by round for Spain*

area sampling design had to be applied. According to this design, first, the population is stratified into H geographical areas. Then, m primary sampling units (i.e. communities) are allocated proportionally to these strata by Cox' controlled rounding method.

In each community, a fixed number of households is selected through a random route procedure. This procedure starts with a random selection of four start addresses from the telephone book. Then, fieldwork personnel (other than the person conducting the interview) follow a random route procedure and collect addresses which are then returned to the fieldwork agency. Within a sampled household, an individual is selected by the last-birthday-method.

The basic sample design did not change over the three rounds under consideration. However, the number of PSUs was constantly increased from 169 (round 1) over 200 (round 2) to 258 in round 3. In addition, also the net sample size was increase from 1 503 (round 1) to 1 806 in round 2 and further to 1 986 in round 3. These changes are graphically depicted in figures 6.69(a) and 6.69(b). As the increase in n_{net} from round 1 to round 2 is over-proportional to the increase in m, the ratio of n_{net} to m, $\bar{b}$, is (slightly) higher in round 2 than in round 1. This is illustrated in figure 6.69(c). As the composition of primary sampling units did not change over the rounds, $\widehat{deff}_c$ can be expected to decrease with $\bar{b}$ from round 1 to round 3.

The distribution of normalized design weights is illustrated in figure 70. Here we can see that the overall distribution hardly changed from round 1 to round 3. There



(a) m (b) n_{net} (c) $\bar{b}$

Figure 69: *Development of m, n_{net} and $\bar{b}$ in France over round 1, 2 and 3*

was, however, a need to truncate 80 design weights in round 1 and one in round 2^{25} to a maximum of 4.0. This is why the distribution in round 1 looks somewhat smoother than in the other rounds. In round 1, the coefficient of variation (cv) of inclusion probabilities at stage 1 (PROB1) is 1.21, at stage 2 (PROB2) 1.17 and at stage 3 (PROB3) 1.16 but the cv of the product of PROB1 and PROB2 is only 1.02. In round 2, however, this observation is even more pronounced. Here, the cv of PROB1 is 2.11, 1.97 for PROB2 and .44 for PROB3. The cv of the product of inclusion probabilities of the first two stages is .12. In round 3, the cv of PROB1 is 2.35, the cv of PROB2 is 1.00 and .43 for PROB3, the cv of PROB1×PROB2 being .11. This indicates that a mayor share of the overall variation in inclusion probabilities is due to variation in PROB3, especially in round 2 and 3.

6.3.3 Poland

In Poland, a dual frame sample design was established in round 1 which has seen some improvements over the rounds. Basically, the population is divided into two domains: the first domain consists of big towns[26], the second domain of all other towns and villages. In the first domain, a simple random sample of individuals is drawn directly from the PESEL frame, a national register of citizens.

In the second domain, a two-stage cluster sample is conducted. At the first stage, m primary sampling units (i.e. small towns and villages) are selected with probability proportional to population size and with replacement. At the second stage, within a PSU, a fixed number of individuals is selected from the PESEL frame by srs. If a PSU at the first stage is selected more than once, o times, say, $o\times$ the fixed number of individuals are selected and treated as an additional, independent PSU. Due to the change in the definition of domains since round 2, the first domain from round 2 onwards contains a greater share of the population (39.49% in round 2 and 39.05%

25 Truncation of design weights is a common method to avoid overemphasized influence of single observations.

26 In round 1, the definition of the first domain included towns of 100 000+ inhabitants. This definition was changed from round 2 onwards where the boundary was lowered to 50 000+ inhabitants.

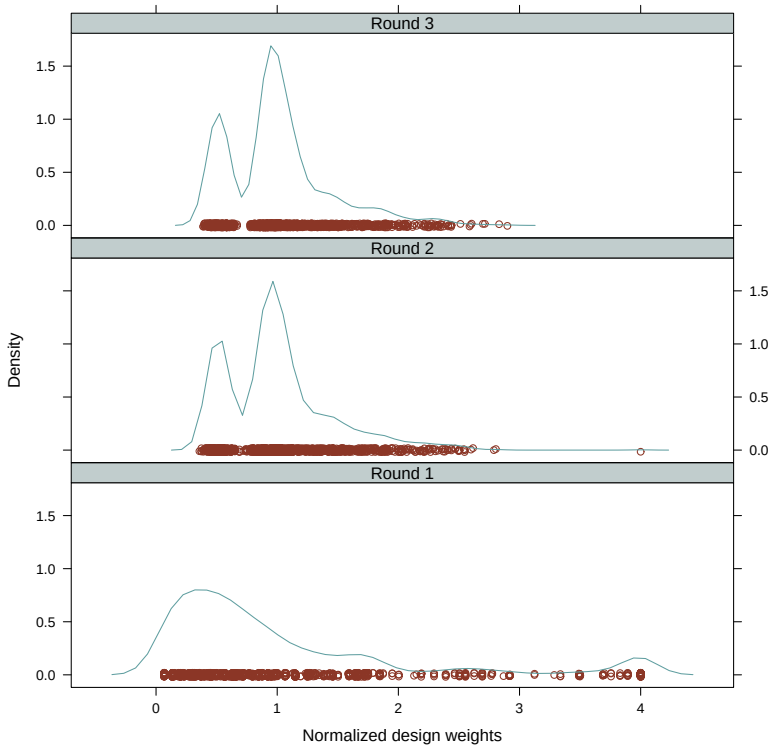


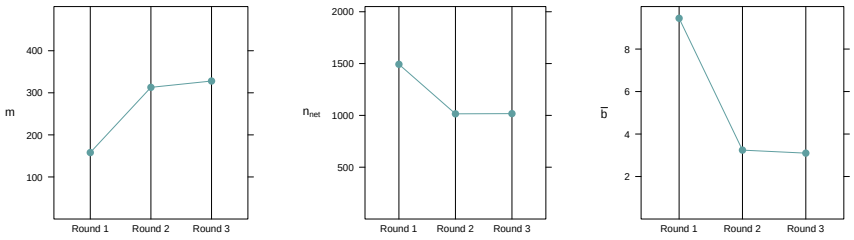
Figure 70: *Grouped density plot of normalized weights by round for France*

in round 3) than under the definition of round 1 (31%). In addition, the number of sampled PSUs in the second domain was increased from round 1 (158) over round 2 (313) to round 3 (328) as illustrated in figure 6.71(a). As $\widehat{deff}$ of a dual-frame sample is a combination of the single design effects within domains (see section 3.1) and $\widehat{deff}_c$ in the first domain is 1, it is easy to see that with a greater share of the sample belonging to the first domain, the overall design effect will decrease. In addition to the increase in m, n_{net} decreased, resulting in an overall decrease in $\bar{b}$, especially from round 1 to round 2. This development is graphically depicted in figures 6.71(b) and 6.71(c).

The distribution of normalized design weights is illustrated in figure 72.

6.3.4 Finland

In Finland, a simple random sample of individuals is drawn from the population register in all three rounds under consideration. In such a case, thinking of sample designs more generally, each respondent can be regarded as a primary sampling unit. Hence, m and n_{net} are equal which is indicated in figures 6.73(a) and 6.73(b). This,



(a) m (b) n_{net} (c) $\bar{b}$

Figure 71: *Development of m, n_{net} and $\bar{b}$ in Poland over round 1, 2 and 3*

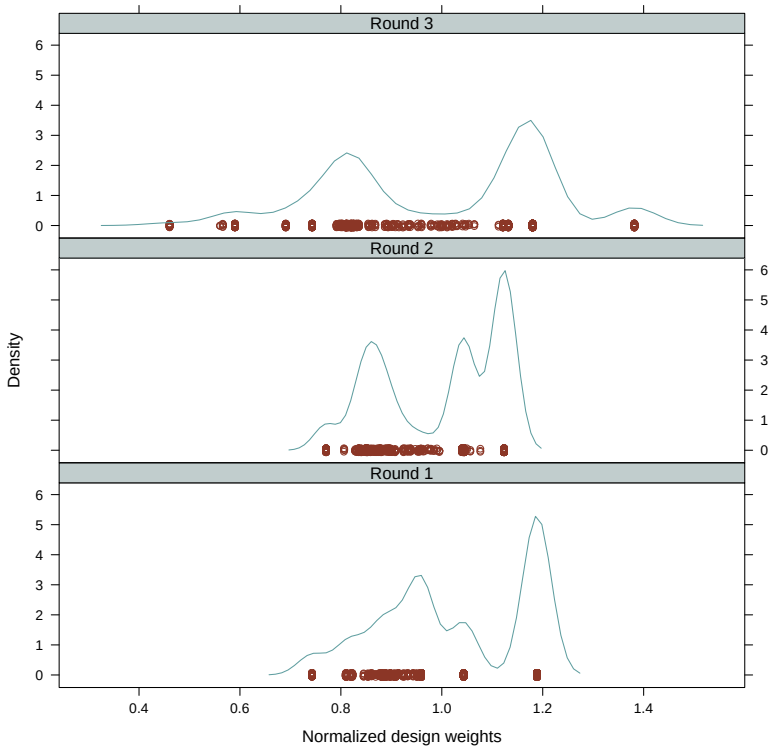
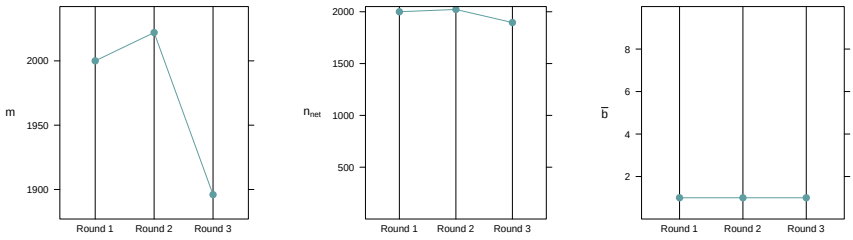


Figure 72: *Grouped density plot of normalized weights by round for Poland*

in turn leads to an average cluster size of $\bar{b}$=1 (see figure 6.73(c)). As individuals are sampled with equal probabilities, $\widehat{deff}_p$ is also 1 and the effective sample size equals n_{net}.

The distribution of normalized design weights is illustrated in figure 74.



(a) m (b) n_{net} (c) $\bar{b}$

Figure 73: *Development of m, n_{net} and $\bar{b}$ in Finland over round 1, 2 and 3*

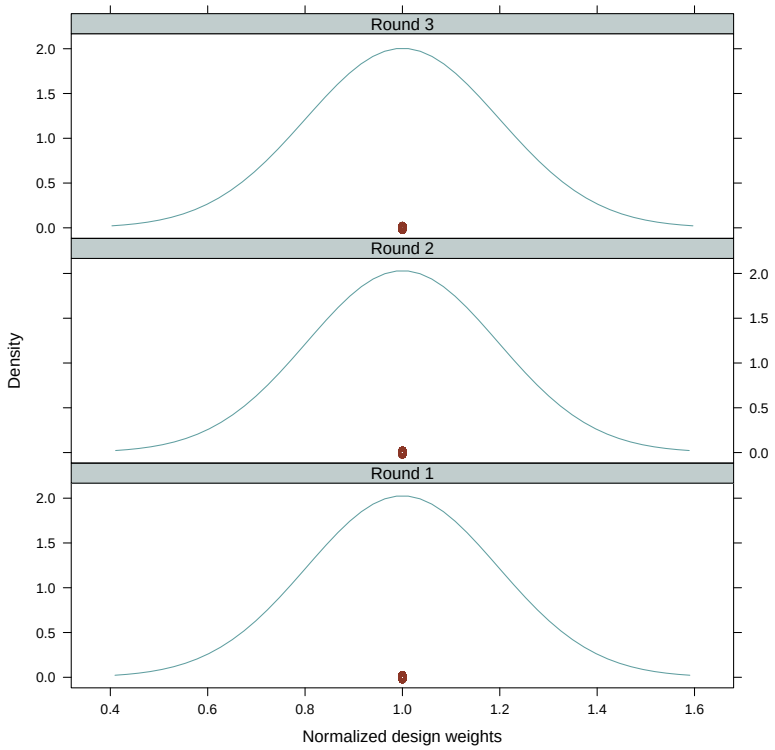


Figure 74: *Grouped density plot of normalized weights by round for Finland*

6.4 Estimation of Design Effects in the ESS

In the ESS, design effects are explicitly used at the planning stage of a sample design (see 6.2). Hence, for the prediction of the required net sample size, based on a fixed effective sample size of n_{eff}=1 500, one needs a good prediction of the expected design

effect in the upcoming round. Here, mainly the quality of $\hat{\rho}$ is of importance since the prediction of ρ for the upcoming round is based on this quantity which is, in turn, estimated with current ESS data.

In a real-world social survey like the ESS, one is confronted with lots of obstacles and irregularities in the data that have an effect on the quality of estimators. Most prominently, non-response can cause problems in many applications. However, since estimation of design effects mainly relies on an estimator's variance, the effect of non-response leading to bias is not considered here. Thus, as far as estimation of design effects and their components (i.e. ρ) is concerned, two important questions to answer are a) what to do with missing values and b) how to handle PSUs with a single observation. Common practice is not to use imputation in the case of missing values. As far as b) is concerned, PSUs with only one observation have to be eliminated since within these clusters, variance estimation will fail.

6.4.1 Estimation of ρ

The usual estimators have also been tested with ESS data. However, unlike in the previous chapter, where estimators could be evaluated against a true population value, with real-world sample survey data from the ESS, such benchmarks do not exist. Hence, as a first step, we shall look at the inter-relation of different estimators by means of correlation tables and plots. Tables 9 to 17 give an overview of the correlation of the different estimators for Likert scaled and continuous items in the ESS core questionnaire. Correlations are estimated separately by ESS for each country[27]. One can easily see that the AOV estimator is highly correlated with most estimators in all countries and rounds[28]. The behaviour of the correlation of AOV with the ML and

Table 9: *Correlations of estimates for Likert and continuous variables in Spain – round 1*

	AOV	F	F2	FR	REML	ML
AOV	1.000					
F	0.938	1.000				
F2	1.000	0.938	1.000			
FR	1.000	0.938	1.000	1.000		
REML	0.994	0.934	0.994	0.994	1.000	
ML	0.994	0.934	0.994	0.994	1.000	1.000

REML estimators is remarkable: In Spain (Tables 9 to 11), where the average cluster size in round 1 was 5.04, 3.65 in round 2 and 3.75 in round 3, the correlation of the AOV with both the ML and REML estimator is highest in the first, lowest in the second round and is at an intermediate level in the third round. This can, in part, be attributed to the instability of the ML and REML methods when cluster sizes are small.

A similar pattern can be observed in France (Tables 12 to 14), although now less pronounced. This is due to the fact that even the smallest average cluster size of $\bar{b}$=7.7 in round 3 is still large enough for the ML and REML estimators to achieve

27 This separation is motivated by the fact that variability of cluster sizes has an impact on some estimators. Hence, simply combining ESS data from different countries can lead to misinterpretation.

28 Estimation of ρ in Poland is based on the clustered part of the sample.

Table 10: *Correlations of estimates for Likert and continuous variables in Spain – round 2*

	AOV	F	F2	FR	REML	ML
AOV	1.000					
F	0.893	1.000				
F2	1.000	0.893	1.000			
FR	1.000	0.893	1.000	1.000		
REML	0.873	0.767	0.873	0.873	1.000	
ML	0.874	0.768	0.874	0.874	1.000	1.000

Table 11: *Correlations of estimates for Likert and continuous variables in Spain – round 3*

	AOV	F	F2	FR	REML	ML
AOV	1.000					
F	0.960	1.000				
F2	1.000	0.960	1.000			
FR	1.000	0.960	1.000	1.000		
REML	0.924	0.881	0.924	0.924	1.000	
ML	0.913	0.873	0.913	0.913	0.976	1.000

Table 12: *Correlations of estimates for Likert and continuous variables in France – round 1*

	AOV	F	F2	FR	REML	ML
AOV	1.000					
F	0.995	1.000				
F2	1.000	0.995	1.000			
FR	1.000	0.995	1.000	1.000		
REML	0.971	0.969	0.971	0.971	1.000	
ML	0.960	0.957	0.960	0.960	0.980	1.000

Table 13: *Correlations of estimates for Likert and continuous variables in France – round 2*

	AOV	F	F2	FR	REML	ML
AOV	1.000					
F	0.790	1.000				
F2	1.000	0.790	1.000			
FR	1.000	0.789	1.000	1.000		
REML	0.971	0.701	0.971	0.971	1.000	
ML	0.972	0.701	0.972	0.972	1.000	1.000

Table 14: *Correlations of estimates for Likert and continuous variables in France – round 3*

	AOV	F	F2	FR	REML	ML
AOV	1.000					
F	0.949	1.000				
F2	1.000	0.949	1.000			
FR	1.000	0.949	1.000	1.000		
REML	0.963	0.860	0.963	0.963	1.000	
ML	0.963	0.861	0.963	0.963	1.000	1.000

good results. Here, correlations between AOV and ML and REML are between .960 (round 1, AOV/ML) and .972 (round 2, AOV/ML).

Turning to the clustered part of the sample in Poland (Tables 15 to 17), we can,

again observe a decrease in correlation between AOV and ML and REML, respectively, as the average cluster size decreases.

Table 15: *Correlations of estimates for Likert and continuous variables in Poland – round 1*

	AOV	F	F2	FR	REML	ML
AOV	1.000					
F	0.955	1.000				
F2	1.000	0.956	1.000			
FR	1.000	0.956	1.000	1.000		
REML	0.980	0.935	0.980	0.980	1.000	
ML	0.981	0.935	0.981	0.981	1.000	1.000

Table 16: *Correlations of estimates for Likert and continuous variables in Poland – round 2*

	AOV	F	F2	FR	REML	ML
AOV	1.000					
F	0.926	1.000				
F2	1.000	0.926	1.000			
FR	1.000	0.926	1.000	1.000		
REML	0.844	0.707	0.844	0.844	1.000	
ML	0.845	0.708	0.845	0.845	1.000	1.000

Table 17: *Correlations of estimates for Likert and continuous variables in Poland – round 3*

	AOV	F	F2	FR	REML	ML
AOV	1.000					
F	0.941	1.000				
F2	1.000	0.941	1.000			
FR	1.000	0.941	1.000	1.000		
REML	0.794	0.743	0.794	0.794	1.000	
ML	0.767	0.717	0.767	0.767	0.956	1.000

In round 1 with $\bar{b}$=9.4, the correlation coefficient of AOV/REML is .980 and .981 for AOV/ML. As the average cluster size goes down to $\bar{b}$=3.2 in round 2, so does the correlation between AOV/REML (.844) and AOV/ML (.845). A further decrease in the average cluster size in round 3 to $\bar{b}$=3.1 causes these correlations for AOV/REML to go down to .794 and for AOV/ML to .767.

The same patterns can also be observed for binary items as can be seen from Tables 25 to 33. The random effects model estimators REML and ML again react sensitively on small average cluster sizes. This holds in Spain, France and especially in Poland where correlations AOV/REML and AOV/REML decrease from .915 (round 1) over .704 (round 2) to .386 (round 3). All correlation tables for binary items can be found in appendix 7.3 (pp. 159).

Turning to differences in $\hat{\rho}$ between the scale type of items, figures 75 and 76 show the distribution of estimates of the different estimators separately for Likert scaled, continuous and binary items. One boxplot in a panel depicts the distribution of all items of a scale type in a specific country at a given round. A comparison of the boxplots within one panel can reveal differences between estimators.

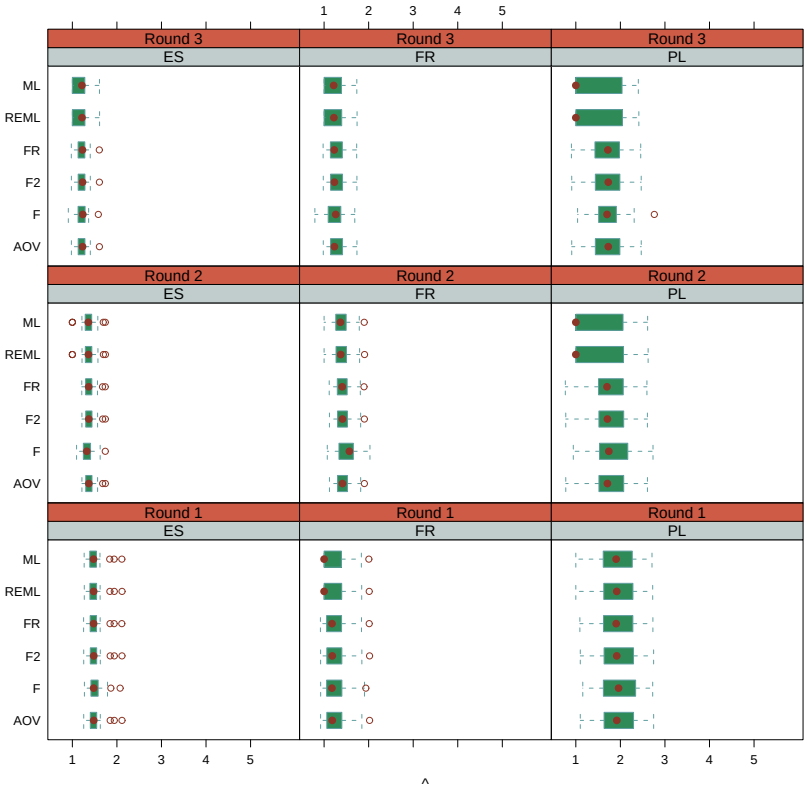


Figure 75: *Grouped boxplots of $\hat{\rho}$ for Likert scaled items*

Comparing boxplots in the lower left panel (i.e. Spain, round 1) of figure 75 we can see hardly any differences in the distribution of the estimates between estimators. The F estimator, however, seems to have a higher variation than all other estimators. In round 2 and 3 we can, however, observe the tendency of the REML and ML estimators to deviate from the classical estimators. This effect, as already mentioned above, has to do with the sensitivity of the variance component estimation to small group (i.e cluster) sizes and is particularly obvious in Spain and Poland in round 3. Here, REML and ML estimators for many items yield values of zero as the estimated variance component of the random effect is estimated zero.

For binary items the same estimators have been investigated as in the simulation studies. Also with dichotomous items we can observe a downwards biased tendency of the REML and ML estimators when average cluster sizes are small (as in Spain and Poland in round 2 and 3). Again, groups of estimators can be found that behave very similarly. For example, AOV, KPR, UBE and FLC show very similar distributions in all panels.

Figures 77 and 78 enable direct comparisons between different estimators for the

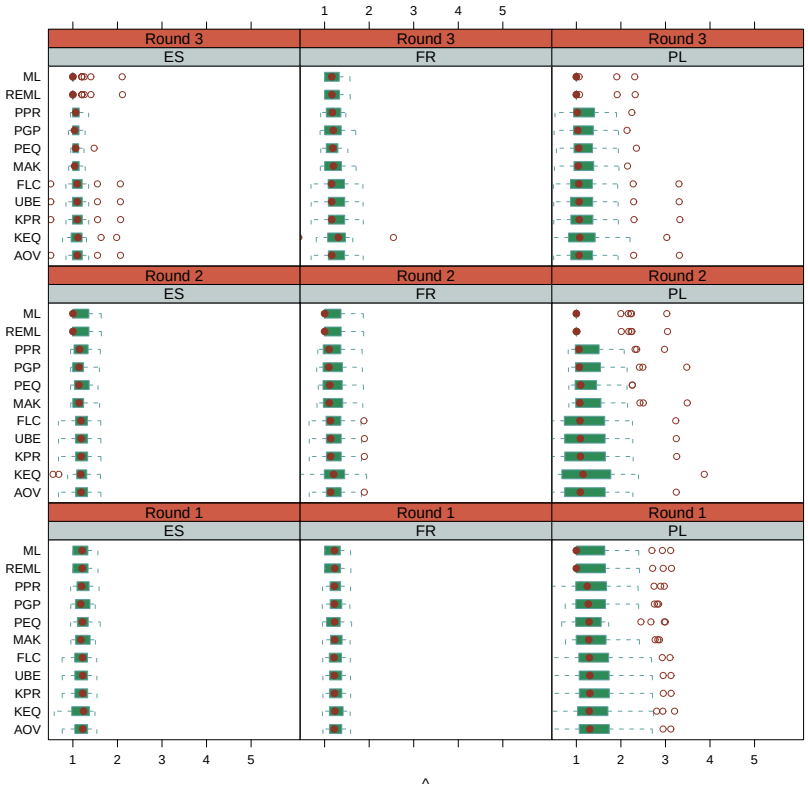


Figure 76: *Grouped boxplots of $\hat{\rho}$ for binary items*

same items. Three items have been selected from the set of all items of each scale type. These items are:

- **Likert** (all listed items measured on 11-point scale)

  LRSCALE political left-right scale,

  STFLIFE overall satisfaction of life and

  HAPPY momentary happiness

- **Continuous**

  HHMMB number of household members,

  YRBRN year of birth of the respondent and

  EDUYRS year of full-time education

- **Binary**

  VOTE indicator for participation at last national election,

  CTZCNTR indicator for citizen of the country and

GNDR gender.

What is more important than the absolute magnitude of the estimates, however, are differences between estimators within a panel.

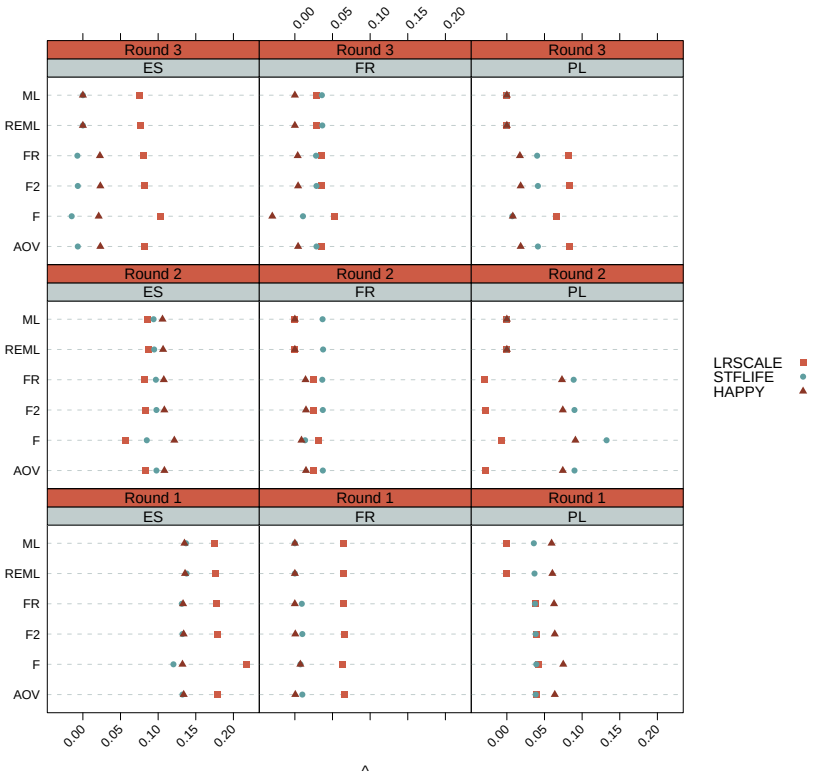


Figure 77: *Grouped dotplots of $\hat{\rho}$ for selected items (Likert)*

With Likert scaled items, we can see that $\hat{\rho}^{(\mathrm{F})}$ breaks ranks in many cases. This effect comes into play when the number of small PSUs is high. In Spain in round 1, for example, the estimate of $\hat{\rho}^{(\mathrm{F})}$ for LRSCALE is considerably higher than the estimates of all other estimators. This is not the case for the item HAPPY where the value of $\hat{\rho}^{(\mathrm{F})}$ is in line with all other estimates. The percentage of extremely small clusters (i.e. clusters of size 2 or 3) is almost 56% for LRSCALE but only 43% for the item HAPPY. The patterns of this example hold in all other cases where $\hat{\rho}^{(\mathrm{F})}$'s estimates differ from those of all other estimators.

The distribution of estimates for binary items is displayed in figure 78. What can be seen at first glance is that $\hat{\rho}^{(\mathrm{MAK})}$ and $\hat{\rho}^{(\mathrm{PGP})}$ have a tendency to yield estimates that differ from those of most other estimators while $\hat{\rho}^{(\mathrm{AOV})}$, $\hat{\rho}^{(\mathrm{KPR})}$, $\hat{\rho}^{(\mathrm{UBE})}$ and $\hat{\rho}^{(\mathrm{PEQ})}$ show very similar values within every panel. Furthermore, due to the definition of $\hat{\rho}^{(\mathrm{REML})}$ and $\hat{\rho}^{(\mathrm{ML})}$ their estimates cannot be negative. This can also be seen in figure 78, most obviously so in Poland where the estimates of GNDR tend to be negative for the

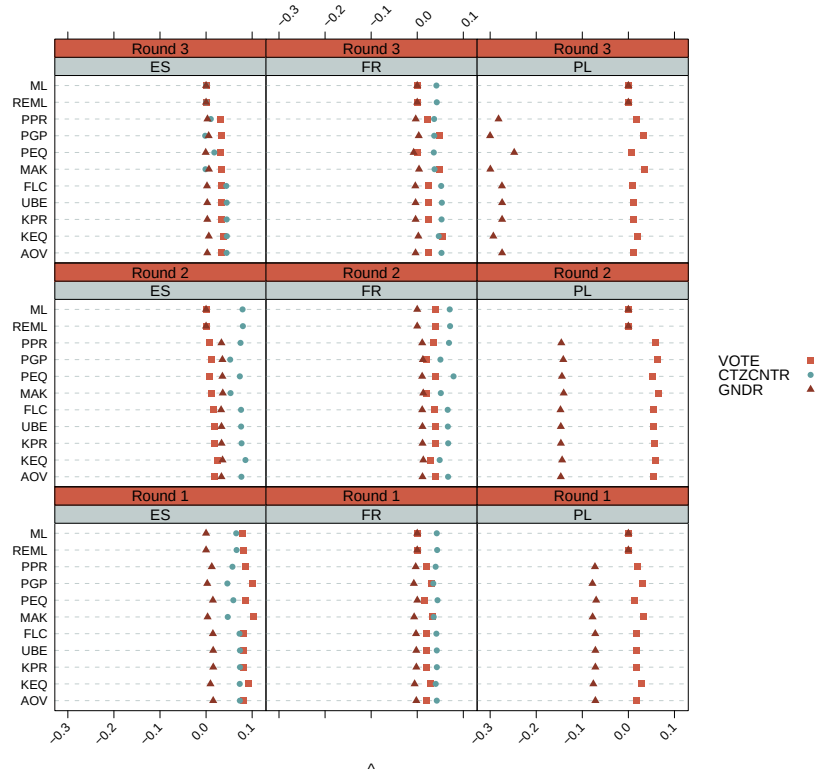


Figure 78: *Grouped dotplots of $\hat{\rho}$ for selected items (Binary)*

classical estimators of ρ but zero for the random effects model based estimators. The fact that in Poland the distribution of estimates for CNZCNTR is missing is caused by the circumstance that all respondents in all rounds have answered this item positively so that there is no variation at all – neither within nor between clusters. Hence most estimators yield NaN due to a division of zero by zero.

6.4.2 Design-based and Model-based Estimation of the Design Effect

In the following, the quality of the estimation of the design effect using the design-based and model-based approach is evaluated. The quality of the estimation of the model-based design effect mainly depends on the quality of $\hat{\rho}$. Thus, everything that has been said about the precision of estimators of ρ will also have an effect on the corresponding model-based estimator of $deff_c$. The design effect due to unequal inclusion probabilities is constant for a sample in a country at a given round of the ESS. The design-based estimators of the design effect, however, directly include both the inflation of variance due to unequal inclusion probabilities as well as the effect of homogeneity within clusters. This is why $\widehat{deff}_{\mathrm{M}}$, and not its components, has to be compared directly to $\widehat{deff}_{\mathrm{D}}$. The distribution of the estimates of appropriate estimators

for selected items of the Likert scale type is illustrated in figure 79. We can see that

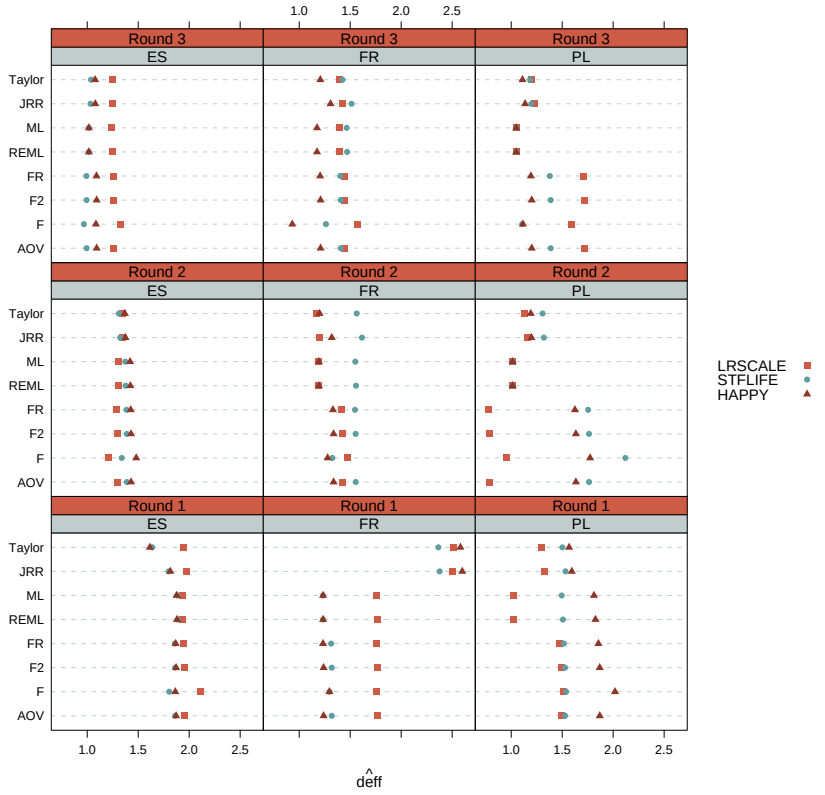


Figure 79: *Grouped dotplots of $\widehat{deff}$ for selected Likert scaled items*

in most panels the model-based and the design-based estimators yield very similar values. However, the sensitivity of $\hat{\rho}^{(F)}$ to small cluster sizes also has an effect on the corresponding model-based estimator of *deff*. Differences between model-based and design-based estimators mainly occur in those cases where design weights show a high variation (i.e. France round 1 and Poland round 2 and 3). This effect is further amplified if the item under consideration is rather skewed (i.e. STFLIFE; not reported here).

The distribution of estimates of selected binary items is shown in figure 80. We can,

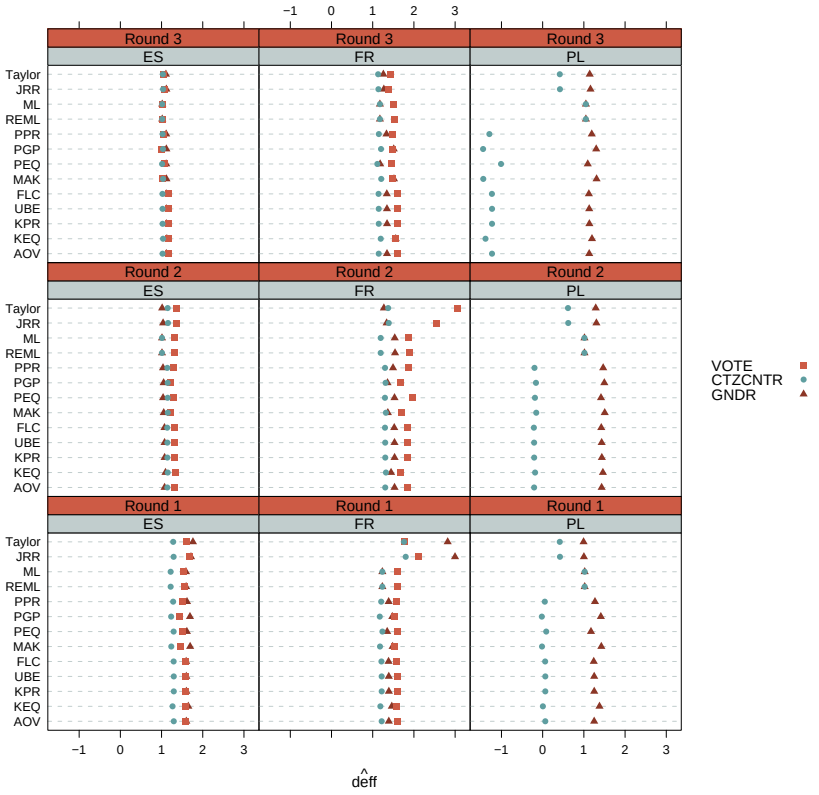


Figure 80: *Grouped dotplots of $\widehat{deff}$ for selected Binary items*

again, observe a tendency of the design-based estimators to yield larger estimates of *deff* when design weights show high variation (i.e. France round 1 and 2) and when the item under consideration is highly skewed (VOTE).

7 Summary

Cross-national sample survey projects like the ESS combine sample data from different countries to enable comparative substantive research. In the ESS, the data are all obtained by fact-to-face interviews but using different sample designs. The bandwidth of sample design used in auch lage-scale multi-national sample survey projects spans from un-stratified simple random sampling (e.g. in Finland) over stratified systematic sampling (e.g. Sweden) to multi-stage cluster sampling (e.g. Portugal or Spain). The reason for these differences in sample designs are differences in the availability of sampling frames among countries. Despite these differences, the specifications of the ESS permit estimators calculated on the basis of one country's sample to be of the same precision as calculated on the basis of another country's sample. One way to make the precision of an estimator independent of the sample design is to conduct samples with comparable effective sample sizes. The concept of the effective sample size, however, incorporates a (model-based) prediction of the design effect. This prediction is based on the estimated value of ρ for certain core variables of the ESS. It is obvious that good (i.e. precise and unbiased) estimates of ρ ensure a high quality of the predicted design effect and hence the predicted required net sample size.

This harmonisation of effective sample sizes is based on a typical design effect and effective sample sizes will vary from item to item since design effects will vary. This is why substantive analyses should either directly use appropriate variance estimators or use the design effect to correct the naive variance estimate.

This thesis has evaluated the quality of diverse estimators of the design effect both derived under the design-based and the model-based perspective. This chapter gives a summary of the empirical findings of the Monte Carlo simulation studies and the applications in the ESS. Section 7.1 recapitalises the basic concept of design effects. In Section 7.2, the use of design effects in the ESS is discussed. Section 7.3 gives an overview of the most important findings that can be derived from the results of Chapter 5.

7.1 Concept of Design Effects

In complex sample surveys, the assumption of independence of observations which underlays the textbook formula for the variance of a point estimator (e.g. the HT estimator of the population mean or total) is often violated. In cluster or multi-stage sampling, for example, elements of the same PSU tend to be more similar to each other than to elements of all other PSUs. This homogeneity, together with the (average) size of PSUs, has a direct influence on the degree to which the naive variance estimator of a point estimator underestimates the true variance of that estimator. The degree of this underestimation is the design effect. Since different items have different levels of homogeneity, design effects will vary from item to item.

If, in addition, the sample design assigns varying inclusion probabilities to ultimate sample elements, for example because the number of elements selected in every PSU

varies, design weighting is necessary. This, in turn, further increases the variance of, for example, the HT estimator of the population mean or a non-parametric measure like the median.

Different approaches to estimation of the design effect rely on different methods to account for both, the effect of variance inflation due to clustering and due to unequal inclusion probabilities. The theoretical considerations together with results of the simulation studies have shown that the design-based approach incorporates both effects directly but suffers from a naive use of the cluster sample data to get an estimate of the variance of a point estimator under srs. The model-based approach, on the other hand, is more flexible in so far as it is possible to estimate both components of *deff* separately. However, it is not guaranteed that there exists a closed-form model-based estimator of *deff* for every estimator under study.

7.2 Use in Complex Sample Surveys

In the ESS, the concept of design effects is used for planning a sample design. To achieve a fixed effective sample size, a typical design effect is predicted. This prediction – if possible – is based on estimates of ρ of selected items of the core questionnaire of the previous round. Researchers are, nevertheless, encouraged to either directly use structural variables such as PSU identifiers in their analyses (if possible, i.e. if a country agrees to make PSU identifiers publicly available) or to use the typical design effect to correct the naive variance estimates. In both cases, estimates should be as precise as possible to avoid additional uncertainty in the results due to low precision in estimated *deff*.

7.3 Estimation of Design Effects and their Components

The answer to the question which estimator of *deff* is favourable depends on some additional parameters like, among other things, the scale type of the study variable, the expected level of ρ, the (average) cluster size or the parameter of interest. Chapter 5 gave in-depth analyses of the quality of estimators of *deff* and of its components under various scenarios. A central finding is that of the model-based estimators, the commonly used classical estimators of ρ like $\hat{\rho}^{(\mathrm{AOV})}$ or $\hat{\rho}^{(\mathrm{F})}$ work very well in scenarios with equal cluster sizes and react with some increase in bias and precision when cluster sizes vary. This is also true for most of the estimators for binary study variables. Here, bias and precision is, however, also influenced by the overall rate of success, π, of the study variable. Almost all of the classical estimators react sensitively if the study variable is heavily skewed. Here, under certain circumstances, $\hat{\rho}^{(\mathrm{REML})}$ and $\hat{\rho}^{(\mathrm{ML})}$ can give better results – mostly when average cluster sizes are small and the parameter of ρ is rather large.

When design and interviewer effects are present at the same time and the variance introduced by interviewer clusters is ignored, ρ_{PSU} may be under- or over estimated – in every case, its estimation is inefficient. However, not all estimators react on such an artefact in the same manner. In many scenarios, $\hat{\rho}^{(\mathrm{ML})}_{\mathrm{PSU}}$ and $\hat{\rho}^{(\mathrm{F2})}_{\mathrm{PSU}}$ tend to be less

biased than all other estimators but in some scenarios quite inefficient. A precise estimation of the share of a certain variance component on the total variance is difficult in all scenarios. However, the decrease in precision is hardly influenced when switching from equal to unequal cluster sizes. Due to the large amount of factors of the Monte Carlo simulation set-ups, for an in-depth evaluation of the quality of different estimators the reader has to be referred to the respective sections.

The patterns found in the results of the Monte Carlo simulation studies can also be observed in results based on data of selected samples of the ESS. Chapter 6 gave an overview of the behaviour of the same estimators that were also under investigation in the simulation studies. Due to the lack of an external criterion, the evaluation of the estimators' quality must follow a different logic. What can be said, however, is that estimators which behave similarly in the artificial environment of the Monte Carlo studies also behave similar in a real-world setting.

Bibliography

Baujat, B., Mahé, C., Sargent, D. J., and Pignon, J.-P. (2001). Heterogeneity in an individual patient data meta-analysis: contribution of random effect survival models. In *9th International Cochrane Colloquium*, Lyon, France.

Bieler, G. S. and Williams, R. L. (1990). Generalized standard error models for proportions in complex surveys. In *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Binder, D. (1996). Linearization methods for single phase and two-phase samples: A cookbook approach. *Survey Methodology*, 22:17–22.

Callens, M. (2006). *Essays on multilevel logistic regression.* Phd thesis, Dept. Applied Economics, K. U. Leuven.

Campbell, M. K., Mollison, J., Steen, N., Grimshaw, J. M., and Eccles, M. (2000). Analysis of cluster randomized trials in primary care: a practical approachfamily practice. *Family Practice*, 17(2):192–196.

Canty, A. J. and Davison, A. C. (1999). Resampling-based variance estimation for labour force surveys. *The Statistician*, 48(3):379–391.

Chantala, K. and Tabor, J. (1999). Strategies to perform a design-based analysis using the add health data. Technical report, Carolina Population Center, University of North Carolina at Chapel Hill.

Choi, J. W. (1987). A direct estimation of intracluster correlation. In *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Choi, J. W. (1989). Variance of intracluster correlation estimator. In *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Clemmer, A. F. and Kalsbeek, W. D. (1984). Estimating intracluster homogeneity in multistage samples. In *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Cochran, W. G. (1977). *Sampling techiques.* Wiley Series in Probability and Mathematical Statistics.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Corbeil, R. R. and Searle, S. R. (1976). Restricted maximum likelihood (reml) estimation of variance components in the mixed model. *Technometrics*, 18(1):31–38.

Cornfield, J. (1951). Modern methods in the sampling of human populations. *American Journal of Public Health*, 41:654–661.

Cummings, W. B. and Gaylor, D. W. (1974). Variance component testing in unbalanced neted designs. *Journal of the American Statistical Association*, 69:765–771.

Davison, A. and Sardy, S. (2007). Resampling variance estimation in surveys with missing data. *Journal of Official Statistics*, 23(3):371–386.

Demnati, A. and Rao, J. N. K. (2002). Linearization variance estimators for survey data. In *Proceedings of the Survey Methods Section*, Proceedings of the Survey Methods Section. SSC Annual Meeting.

Deville, J. (1999). Variance estimation for complex statistics and estimators: Linearization and residual techniques. *Survey Methodology*, 25:193–203.

Donner, A. (1986). A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. *International Statistical Review*, 54(1):67–82.

ESS (2005a). European social survey, round 3: Specification for participating countries. Specification, European Social Survey.

ESS (2005b). Sampling for the european social survey round iii: Principles and requirements. Specification, European Social Survey.

Faraway, J. J. (2006). *Extending the Linear Model with R*. Chapman & Hall.

Fields, J. M. (1970). The cluster sample: A neglected data source. *The Public Opinion Quarterly*, 34(4):593–603.

Fleiss, J. L. and Cuzick, J. (1979). The reliability of dichotomous judgments: Unequal numbers of judges per subject. *Applied Psychological Measurement*, 3(4):537–542.

Gabler, S., Ganninger, M., and Lahiri, P. (2010). Randomization-based versus model-based design effects: a comparison. in submission.

Gabler, S. and Häder, S. (2000). *Über Design Effekte*, chapter 4, pages 73–97. http://193.175.239.100/Publikationen/Aufsaetze/ZUMA/Festschrift

Gabler, S., Häder, S., and Lahiri, P. (1999). A model based justification of Kish's formula for design effects for weighting and clustering. *Survey Methodology*, 25(1):105–106.

Gabler, S., Häder, S., and Lynn, P. (2006). Design effects for multiple design surveys. *Survey Methodology*, 32(1):115–120.

Gabler, S. and Lahiri, P. (2009). On the definition and interpretation of the interviewer variability for a complex sampling design. *Survey Methodology*, 35(1):85–99. forthcoming.

Gibbons, R. D. and Hedeker, D. (1994). Application of random-effects probit regression models. *Journal of Consulting and Clinical Psychology*, 62(2):285–296.

Gibbons, R. D. and Hedeker, D. (1997). Random effects probit and logistic regression models for three-level data. *Biometrics*, 53(4):1527–1537.

Gilks, W. R., Wang, C. C., and Coursaget, B. Yvonnet, P. (1993). Random-effects models for longitudinal data using gibbs sampling. *Biometrics*, 49(2):441–453.

Goldstein, H. and Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society*, 159(3):505–513.

Gonzalez, E. J. and Foy, P. (2004). *Estimation of Sampling Variability, Design Effects, and Effective Sample Sizes*, chapter 5, pages 82–200. TIMSS & PIRLS International Study Center.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, 72(358):320–338.

Harville, D. A. (2004). Making reml computationally feasible for large data sets: Use of the gibbs sampler. *Journal of statistical computation and simulation*, 74(2):135–153.

Häder, S., Laaksonen, S., and Lynn, P. (2007). *ESS Round 2 2004/2005 Technical Report*, chapter THE SAMPLE. ESS.

Hedeker, D. and Mermelstein, R. J. (1996). Application of random-effects regression models in relapse research. *Addiction*, 91 (Suppl.):211–229.

Irwin, J. O. (1946). On the interpretation of the within and between class analysis of variance when the intra class correlation is negative. *Journal of the Royal Statistical Society*, 109(2):157–158.

Katz, J., Carey, V. J., Zeger, S. L., and Sommer, A. (1993). Estimation of design effects and diarrhea clustering within households and villages. *American Journal of Epidemiology*, 138(11):994–1006.

Kish, L. (1965). *Survey Sampling*. John Wiley & Sons.

Kish, L. (1987). Weighting in deft^2. *The Survey Statistician*, 17(1):26–30.

Kish, L. (1989). Deffs: Why, when and how? a review. In *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Kish, L. (1994). Multipopulation survey designs: Five types with seven shared aspects. *International Statistical Review*, 62:167–186.

Kish, L. (1995). Methods for design effects. *Journal of Official Statistics*, 11(1):55–77.

Kleinman, J. C. (1973). Proportions with extraneous variance: Single and independent samples. *Journal of the American Statistical Association*, 68(341):46–54.

Lago, J. A., Massey, J., Ezzati, T., Johnson, C., and Fullwood, R. (1987a). Evaluation of design effects for the mexican american portion of hispanic health and nutrition examination survey. In *Proceedings of the Survey Research Methods Section, American Statistical Association*.

Lago, J. A., Massey, J., Ezzati, T., Johnson, C., and Fulwood, R. (1987b). Evaluation of design effects for the mexican american portion of the hispanic health and nutrition examination survey. In *Proceedings of the Survey Research Methods Section*.

Lange, N. and Ryan, L. (1989). Assessing normality in random effects models. *The Annals of Statistics*, 17(2):624–642.

Lehtonen, R., Djerf, K., Härkänen, T., and Laiho, J. (2002). Design-based and model-based methods in analyzing complex health survey data: A case study. In *Proceedings of Statistics Canada Symposium*.

Lehtonen, R. and Pahkinen, E. (2004). *Practical Methods for Design and Analysis of Complex Surveys*. Wiley.

Lemeshow, S., Letenneur, L., Dartigues, J.-F., Lafont, S., Orgogozo, J.-M., and Commenges, D. (1998). Illustration of analysis taking into account complex survey considerations: The association between wine consumption and dementia in the paquid study. *American Journal of Epidemiology*, 148(3):298–306.

Liu, Q. and Pierce, Donald, A. (1994). A note on gauss-hermite quadrature. *Biometrika*, 81(3):624–629.

Lohr, S. (1999). *Sampling: Design and Analysis*. Duxbury Press.

Lynn, P. and Gabler, S. (2005). Approximations to b* in the prediction of design effects due to clustering. *Survey Methodology*, 31(2).

Mak, T. K. (1988). Analysing intraclass correlations for dichotomous variables. *Applied Statistics*, 37(3):344–352.

Martinez, B. and Brogan, D. (1984). A comparison of methods for estimating the intraclass correlation of blood pressure. In *Proceedings of the Survey Research Methods Section, American Statistical Association*.

McCullagh, P. and Nelder, J. (1983). *Generalized Linear Models*, volume 1 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, London, New York.

McCulloch, Charles, E. (1994). Maximum likelihood variance component estimation for binary data. *Journal of the American Statistical Association*, 89(425):330–335.

McGraw, K. O. and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1(1):30–46.

McKean, J. and Schrader, R. (1984). A comparison of methods for studentizing the sample median. *Communications in Statistics*, 13:751–773.

Mentre, F., Mallet, A., and Baccar, D. (1997). Optimal design in random-effects regression models. *Biometrika*, 84(2):429–442.

Münnich, R. (2003a). Data quality in complex surveys. In *Bulletin of the International Statistical Institute, 54th Session*, volume LX, pages 164 – 165.

Münnich, R. (2003b). On the optimal design in stratified regression estimation. *Allgemeines Statistischen Archiv*, 87(1):25–38.

Münnich, R. (2004). Varianzschätzung im Deutschen Mikrozensus und dessen Bedeutung für die Qualität der Angabewerte. In *Statistische Analysen: Kolloquium 2002 in Baden-Württemberg*, volume 2, pages 15–23.

Münnich, R. (2008). Varianzschätzung in komplexen erhebungen. *AUSTRIAN JOURNAL OF STATISTICS*, 37(3&4):319–334.

Münnich, R. and Rässler, S. (2004). Variance estimation under multiple imputation. In *Proceedings of the Q2004 conference*.

Paul, S. R. (1990). Maximum-likelihood estimation of intraclass correlation in the analysis of family data: Estimating equation approach. *Biometrika*, 77(3):549–555.

Paul, S. R., Saha, K. K., and Balasooriya, U. (2003). An empirical investigation of different operation characteristics of several estimators of the intraclass correlation in the analysis of binary data. *Journal of Statistical Computation & Simulation*, 73(7):507–523.

Rabe-Hesketh, S. and Skrondal, A. (2002). Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2:1–21.

Rao, J. and Wu, C. (1988). Resampling inference with complex survey data. *Journal of the American Statistical Association*, 83(401):231–241.

Rao, J. N. K. and Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86:403–415.

Ridout, M. S., Demétrio, C. G. B., and Firth, D. (1999). Estimating intraclass correlation for binary data. *Biometrics*, 55:137–148.

Rodríguez, G. and Elo, I. (2003). Intra-class correlation in random-effects models for binary data. *The Stata Journal*, 3(1):32–46.

Rodriguez, G. and Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society*, 158(1):73–89.

Rothery, P. (1979). A nonparametric measure of intraclass correlation. *Biometrika*, 66(3):629–639.

Rowe, A. K., Lama, M., Onikpo, F., and Deming, M. S. (2002). Design effects and intraclass correlation coefficients from a health facility cluster survey in benin. *International Journal of Quality in Health Care*, 14(6):521–523.

Santos, D. and Berridge, D. (1999). A random effects model for repeated ordinal responses with application to breast cancer data. In *Bulletin of the International Statistical Institute*.

Schnell, R., Hill, P., and Esser, E. (1999). *Methoden der empirischen Sozialforschung*. Oldenbourg.

Schnell, R. and Kreuter, F. (2005). Separating interviewer and sampling-point effects. *Journal of Official Statistics*, 21(3):389–410.

Selhub, J., Jacques, P. F., Rosenberg, I. H., Rogers, G., Bowman, B. A., Gunter, E. W., Wright, J. D., and Johnson, C. L. (1999). Serum total homocysteine concentrations in the third national health and nutrition examination survey (1991-1994): Population reference ranges and contribution of vitamin status to high serum concentrations. *Annals of Internal Medicine*, 131(5):331–339.

Shrout, P. E. and Fleiss, J. L. (1979). Intraclass correlations : Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Siddiqui, O., Hedeker, D., Flay, B. R., and Hu, F. B. (1996). Intraclass correlation estimates in a school-based smoking prevention study. *American Journal of Epidemiology*, 144(4):425–433.

Skinner, C. J. (1986). Design effects of two-stage sampling. *Journal of the Royal Statistical Society*, 48:89–99.

Särndal, C.-E., B., S., and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer-Verlag.

Stiratelli, R., Laird, N., and Ware, J. H. (1984). Random-effects models for serial observations with binary response. *Biometrics*, 40(4):961–971.

Tamura, R. N. and Young, S. S. (1987). A stabilized moment estimator for the beta-binomial distribution. *Biomelrics*, 43(4):813–824.

Taylor, B. (1969). *A Source Book in Mathematics 1200-1800*, chapter Methodus Incrementorum Directa et Inversa [Direct and Reverse Methods of Incrementation], pages 329–332. Harvard University Press, Cambridge, Massachusetts.

Thomas, K. F., Earner, D. A., and Fay, R. (1983). Intraclass correlations using a sample of 1980 census data. In *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Valliant, R., Dorfman, A. H., and Royall, R. M. (2000). *Finite Population Sampling and Inference.* WIley.

Verma, V., Scott, C., and O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the world fertility survey. *Journal of the Royal Statistical Society*, 143(4):431–473.

Walsh, J. E. (1947). Concerning the effects of intraclass correlation on certain significance tests. *Annals of Mathematical Statistics*, 18:88–96.

Williams, D. A. (1982). Extra-binomial variation in logistic linear models. *Applied Statistics*, 31:144–148.

Wolter, K. M. (1985). *Introduction to Variance Estimation.* Springer Series in Statistics. Springer-Verlag.

Yamamoto, E. and Yanagimoto, T. (1992). Moment estimators for the beta-binomial distribution. *Journal of Applied Statistics*, 19(2):273–283.

Yeo, D., Mantel, H., and Liu, T.-P. (1999). Bootstrap variance estimation for the national population health survey. In *Proceedings of the Survey Research Methods Section, American Statistical Association.*

Zeger, S. L., Liang, K.-Y., and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, 44(4):1049–1060.

Zou, G. and Donner, A. (2004). Confidence intervals estimation of the intraclass correlation coefficient for binary outcome data. *Biometrics*, 60:807–811.

Appendix

Excursus: Generalized Linear Models
Excursus: Random Effect Models
Tables

Excursus: Generalized Linear Models

This excourse briefly explores the generalized linear model (GLM). This is done for the sake of a better understanding of the random effect models which underlay the estimators of ρ introduced in Chapter 4. Random effect models are also briefly delineated in the next chapter on pages 155 to 158.

The GLM is a generalization of the well known linear model (LM). Through the use of a link function, GLMs are very flexible and allow modelling relationships between predictors and responses from a large family of distributions. The first section of this chapter gives a brief introduction into the linear model to lay the foundations for delineation of GLMs in Section 7.3.

The Linear Model

The basic linear model allows for modeling simple relationships between two or more variables. The dependent variable, however, is assumed to be at least of interval scale. The independent variable(s), on the other hand, may be continuous, discrete or categorical. Analyses based on linear models are more commonly referred to as regression analyses. Depending on the number of predictors, p, we shall speak of *simple regression* if $p = 1$ and of *multiple regression* when $p > 1$. Put most general, a linear model is of the form

$$Y = f(X) + \epsilon \quad , \tag{1}$$

where Y is the response to be modeled through f, an unknown function on X, a predictor variable. The error made through this representation is captured by $\epsilon \sim N(0, \sigma)$. In the most simple case X is a $n \times 1$ vector, where n is the number of cases for which a prediction on Y is to be made. Unless f is undefined, this representation, of course leaves infinite possibilities of estimating its parameters. Under the restriction of f to be a linear function, one possible modification of 1 is

$$Y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X + \epsilon \quad , \tag{2}$$

where $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are now the unknown parameters to be estimated. The above formulations would refer to a simple regression model with only one predictor variable. A generalization of the model above would allow X to be of any dimension, $p > 1$. Thus, $\mathbf{X}$ now is a $n \times p$ matrix defined as

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}.$$

The model may now be written as

$$Y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \boldsymbol{X}_1 + \boldsymbol{\beta}_2 \boldsymbol{X}_2 + \ldots + \boldsymbol{\beta}_p \boldsymbol{X}_p + \epsilon \tag{3}$$

or simple in matrix notation as

$$Y = \mathbf{X}\boldsymbol{\beta} + \epsilon \qquad , \tag{4}$$

where $\boldsymbol{\beta}$ now is a $p+1$ column vector defined as $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)^T$ and $\mathbf{X}$ now is a $n \times p+1$ matrix defined as

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ & & & & \\ 1 & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

where the ones in the first column indicate the multiplier for the intercept term.

Definition of a GLM

A GLM specifies a relation between a *linear predictor*, η, and the expected value of a *dependent variable* or *response* from the exponential family, $\theta = \mu$, through a *link function*, g^{-1}, and is hence given by:

$$E(\boldsymbol{Y}) = \mu = g^{-1}(\boldsymbol{X}\beta) = g^{-1}(\mu) \qquad . \tag{5}$$

Exponential Family of the Response

A density function, Y, is a member of the exponential family if and only if its distribution function is of the form

$$f(y|\theta,\phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right] \qquad , \tag{6}$$

where θ and ϕ are the *canonical* (location) and the *dispersion parameters*, respectively. The functions $a(\cdot)$, $b(\cdot)$, and $c(\cdot)$ are real valued functions and may be chosen appropriately to account for a certain type of distribution.

Many important distribution functions can be expressed through (6). Choosing, for example, $\theta = \mu$, $\phi = \sigma^2$, $a(\phi) = \phi$, $b(\theta) = \frac{\theta^2}{2}$, and $c(y,\phi) = -\frac{\left(\frac{y^2}{\phi} + \log(2\pi\phi)\right)}{2}$, the *normal distribution* results from (6) as

$$f(y|\theta,\phi) = \exp\left[\frac{y\mu - \mu^2/2}{\sigma^2} - \frac{1}{2}\left(\frac{y^2}{\sigma^2} + \log(2\pi\sigma^2)\right)\right].$$

The *binomial distribution* can be expressed through (6) by choosing $\theta = \log\frac{\mu}{1-\mu}$, $b(\theta) = -n\log(1-\mu) = n\log(1+\exp\theta)$, and $c(y,\phi) = \log\binom{n}{k}$, and setting $\phi = 0$ and $a(\phi) = 0$, which results in

$$f(y|\theta,\phi) = \exp\left[y\log\frac{\mu}{1-\mu} + n\log(1-\mu) + \log\binom{n}{k}\right].$$

Of course, the exponential family distributions are characterized by a mean, $\mathrm{E}(Y)$, and a variance, $\mathrm{Var}(Y)$. These are, respectively, given by

$$\begin{aligned} \mathrm{E}(Y) &= \mu = b'(\theta) \\ \mathrm{Var}(Y) &= b''(\theta)a(\phi) \end{aligned}$$

Note that the expected value of Y is a function of θ only, whereas, in general, the variance also takes into account ϕ, the scale parameter. The variance of the binomial distribution, for example, is a case where the variance is a function of the mean. The Gaussian distribution, on the other hand, has variance function that does not depend on the mean, so $b''(\theta) = 1$.

The Linear Predictor

The prediction of the response is by means of a linear predictor, η, which is of the form

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \boldsymbol{x}^T \boldsymbol{\beta},$$

where p is the number of predictors, x are predictors of the model and the β vectors are the vectors of unknown coefficients which are subject to estimation. The right-hand side of the above equation is the equivalent matrix expression where $\boldsymbol{x}$ is a $n \times p$ matrix of predictors and $\boldsymbol{\beta}$ is the $n \times p + 1$ matrix of coefficients.

The Link Function

A *link function*, g, relates the linear predictor to the distribution function through its mean, $\mathrm{E}(Y) = \mu$. It thus accounts for the non-linearity of the relationship between the response and elements of the linear predictor. Thus, we can write

$$\eta = g(\mu) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \boldsymbol{x}^T \boldsymbol{\beta}.$$

Hence, the mean response can be expressed as

$$\mathrm{E}(Y) = g^{-1}(\mu) \qquad .$$

This poses the question, which link functions are suitable for a given type of response. Several link functions have been proposed (McCullagh and Nelder, 1983). It is convenient to choose g in a way that $\eta = g(\mu) = \theta$. In that case, g is called *canonical link*. The most important canonical link functions are summarized in table 18.

Table 18: *Link, mean, and variance functions for selected members of the exponential family; from Faraway (2006, 117)*

Distribution	Link Function	Mean Function	Variance Function
Normal	$\eta = \mu$	$\mu = \eta$	1
Poisson	$\eta = \log(\mu)$	$\mu = \exp(\eta)$	μ
Binomial	$\eta = \log\left(\frac{\mu}{1-\mu}\right)$	$\mu = \frac{\exp(\eta)}{1+\exp(\eta)}$	$\mu(1-\mu)$
Gamma	$\eta = \mu^{-1}$	$\mu = \eta^{-1}$	μ^2
Inv. Gaussian	$\eta = \mu^{-2}$	$\mu = \eta^{-1/2}$	μ^3

Estimation

If values of $\boldsymbol{Y}$, Y_i with $i = 1, \ldots, n$ say, can be assumed to be iid[29] the log-likelihood for observation i is

$$\log L\left(\theta, \Phi, y_i\right) = w_i \left(\frac{y_i \theta_i - b(\theta_i)}{\Phi} \right) + c(y_i, \Phi) \qquad , \tag{7}$$

where w_i are known weights (Faraway, 2006, 116). Hence the overall log-likelihood is

$$\sum_{i=1}^{n} \log L\left(\theta_i, \Phi, y_i\right) \qquad . \tag{8}$$

The log-likelihood can only be maximised analytically in the iid case when the response is Gaussian. In all other cases iterative numerical procedures must be applied. The most prominent of these procedures is Newton-Raphson with Fisher scoring and Iteratively Reweighed Least Squares (Faraway, 2006, pp. 117).

29 iid denotes *identically independent distributed* and implies that the y values are independent realizations of the respective distributions

Excursus: Random Effect Models

When the iid assumption in a GLM is violated, for example because population elements are grouped together according to a certain structure which has influence on the study variable, *random effects models* (REM) offer one possibility to account for this underlying structure. The random effects in such a model have a variance, accounting for the fact that, for example, membership in clusters can differ in strength of influence. This variance component, as one part of the total variance, is a model parameter which is used for the estimation of ρ (see Sections 4.2.2 and 4.3.4).

This chapter first gives an overview of the basic concept behind random effect models. In Section 7.3 a basic model is defined and its applicability to the problem at hand is discussed. A more complex, *hierarchical* (or *nested*) model is presented in Section 7.3 and estimation techniques are discussed in Section 7.3.

Overview

Random effect models are a natural extension to models of the GLM class presented in the previous chapter. They have lots of their properties and can thus be treated and interpreted in similar ways. Besides their application within the context of this thesis, REMs receive prominent attention in biology (Stiratelli et al., 1984; Mentre et al., 1997), medicine and health research (Hedeker and Mermelstein, 1996; Gibbons and Hedeker, 1997; Santos and Berridge, 1999; Lange and Ryan, 1989; Baujat et al., 2001), as well as in the empirical social sciences. As an extension, random effect models are also used when grouping is assumed to arise from serial observations on the same subject (Stiratelli et al., 1984; Gibbons and Hedeker, 1994,9; Gilks et al., 1993; Zeger et al., 1988; Lange and Ryan, 1989).

In a REM, a portion of the variation in the dependent variable can be attributed to group membership. In a *two-way model*, each ultimate sample element (i.e. respondent) is a member of exactly one group. The two *levels* of the model correspond to the respondent (level one) and the random effect (level 2). As a rule, annotation of levels is bottom-up, from smaller to larger grouping elements. It is, however, possible that there exist multiple groups – either aside (*crossed*) or *nested* within each other. For example, interviewer clusters can cross geographic clusters. The model assumes that there is no unique direction into which group membership influences the variable under study – the effects are random with an expected value zero. However, the effect on elements of the same group points into the same direction, i.e. "observations of the same group will be dependent" (Callens, 2006, 2).

From a technical point of view, estimation of the distribution of random effects receives prominent attention. Depending on the structure of the dependent variable, different estimation methods are now implemented in the most wide-spread software packages. They can be classified into maximum-likelihood, quasi-likelihood methods and methods of numerical integration. Comparisons of their effectiveness and convergence behaviour (Callens, 2006; Rodriguez and Goldman, 1995; Goldstein and Rasbash, 1996) form the main line of empirical research in this field.

The Basic Model

Turning to model formulation, many of the definitions given in earlier chapters on GLMs still hold. What extends a GLM to a REM is the inclusion of a random intercept or a random slope or the combination of the two. The most basic random effects model is the one-way ANOVA model:

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij} \qquad i = 1, \ldots, m \quad j = 1, \ldots, n \tag{9}$$

where α represent the random effects (here: intercepts) with m levels, ϵ_{ij} is a random error term and y_{ij} is the outcome of the jth ultimate sample element given the ith level of the random effect. The expected value of both α and ϵ is zero but variances are σ^2_α and σ^2_ϵ, respectively. They are assumed to be $N(0, \sigma^2_\alpha)$ and $N(0, \sigma^2_\epsilon)$; ϵ are assumed to be independent of α (Cummings and Gaylor, 1974, 765). This basic model applies in situations where two-stage cluster sampling is applied and random effects are represented by PSUs. With this type of model, the intraclass correlation coefficient is generally defined as in Sections 4.2.2 and 4.3.4 as

$$\rho = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_\epsilon}.$$

In the case of a binary outcome y_{ij} represents an unobserved continuous variable and we observe only $z_{ij} = I_{y_{ij}} > 0$, i.e. an indicator which is one if y_{ij} exceeds a threshold of 0 and zero otherwise (McCulloch, 1994, 330). We can then treat the model as a GLM, specify an appropriate link function (see Table 18) and add the random effects to the model formula. For the estimation of ρ, however, we are only interested in the estimation of the variance components due to the random effects.

A Hierarchical Model

The above model can be extended quite easily by adding further random effects on a higher or lower level. The most obvious extension is to introduce a further random effect which is *nested* within the first one. This can be conceptualized by use of the two-way ANOVA model:

$$y_{ikj} = \mu + \alpha_i + \beta_{ik} + \epsilon_{ikj} \qquad i = 1, \ldots, m \quad k = 1, \ldots, K, \quad j = 1, \ldots, n \tag{10}$$

with σ^2_α, σ^2_β and σ^2_ϵ being the variances of the two random effects, α and β, and the micro-level error term. The nested random effects (β) are assumed to be balanced within the higher level random effects (α) with $k_i = k$ β-clusters in the ith α-cluster. The intraclass correlation coefficient of the α-level random effect is given by

$$\rho_\alpha = \frac{\sigma^2_\alpha}{\sigma^2_\alpha + \sigma^2_\beta + \sigma^2_\epsilon} \tag{11}$$

and the β-level intraclass correlation coefficient is

$$\rho_\beta = \frac{\sigma_\beta^2}{\sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\epsilon^2} \quad . \tag{12}$$

A two-level random effects model can be useful to account for a nesting of interviewers within geographical clusters or vice versa (see Section 5.7).

Estimation

The most intuitive and direct way to estimate the one-way model and to extract the variance components is to use ANOVA techniques. Leaving out fixed effects, the estimator for σ_α^2 is given by

$$\hat{\sigma}^2_{\alpha,(AOV)} = \frac{\frac{SSB}{(m-1)} - \hat{\sigma}^2_\epsilon}{n} = \frac{MSB - MSW}{n}$$

and the estimator of σ_ϵ^2 is

$$\hat{\sigma}^2_{\epsilon,(AOV)} = \frac{SSW}{m(n-1)} = MSW.$$

The quantities MSW and MSB can be obtained from the usual ANOVA table. This method, however, is based on the assumption that the data are perfectly balanced which will, most likely, not be the case in a real-world sample survey. Besides this one, ANOVA estimation has two further drawbacks (Faraway, 2006, pp. 154): 2.) The errors are assumed to follow a Gaussian distribution. 3.) For more complex models, the estimation becomes cumbersome.

An alternative approach is maximum likelihood estimation (MLE). Here, the variance components, $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_\epsilon^2$, are fit such that the log-likelihood of the parameters given the data is maximised. With MLE, however, one must assume a distribution of the errors. Usually this distribution is assumed to be Gaussian, but can in fact be any other distribution of the exponential family. The random effects model estimator of ρ using ML techniques to estimate the variance components of the model, $\hat{\rho}^{(\mathrm{ML})}$, is obtained by fitting the study variable on the clusters as grouping factors, α, in the model. Then, the estimated variance components $\hat{\sigma}_\alpha$ and $\hat{\sigma}_\epsilon$ are substituted into formula 4.5.

Estimation can be performed with any standard statistical software. In R, however, the random effects model can be ML estimated by use of the lmer() function from the package lme4[30]. The ML estimation method, however, tends to yield downward biased estimates (Harville, 1977, 325) which will lead to underestimation of ρ.

With REML estimation this bias is reduced. Instead of estimating the parameter

30 For the function to estimate the model my ML one must specify the option REML=FALSE.

with the full vector of observations, estimation is based on a decomposition of $n-1$ linearly independent[31] "error contrasts" (see Harville, 1977, 325). Numerical methods are used to maximise the likelihood function. The general procedure is described in Corbeil and Searle (1976). Recently, an improved method based on the GIBBS sampler has been proposed by Harville (2004).

A third estimation method based on REML estimation and its numerical procedures is Laplace approximation. With REML estimation, one has to iteratively evaluate the Gauss-Hermite Liu and Pierce (1994) approximation[32] to the density under the integral derived from the log-likelihood function, a term of the form $\sum_{i=1}^{q} f(x_i)\, e^{x_i^2}\, w_i$ with q points of approximation and weights w_i. The number of points at which the approximation is evaluated can vary – a higher number of points giving higher precision at the cost of computation time. With Laplace approximation the density is approximated at only one such point each time.

In the case of a binary response, variance component estimation is a bit more cumbersome but follows the same principles as in the case of continuous responses. A detailed discussion of these methods cannot be given within the scope of this thesis but McCulloch (1994) gives a comprehensive overview of ML and REML estimation for binary outcome data using the EM algorithm.

31 More generally, $n-p$ where p is the number of fixed effects in the model.

32 This procedure is often referred to as *Adaptive Gaussian Quadrature* or simply *AGQ* (see Rabe-Hesketh and Skrondal, 2002, pp. 5).

Tables

Table 19: *Factors of the simulation study with Gaussian study variables*

no.	cluster size type	m	ρ	scale
1	equal	150	0.01	Gaussian
2	equal	150	0.02	Gaussian
3	equal	150	0.03	Gaussian
4	equal	150	0.04	Gaussian
5	equal	150	0.05	Gaussian
6	equal	150	0.10	Gaussian
7	equal	150	0.15	Gaussian
8	equal	150	0.20	Gaussian
9	equal	300	0.01	Gaussian
10	equal	300	0.02	Gaussian
11	equal	300	0.03	Gaussian
12	equal	300	0.04	Gaussian
13	equal	300	0.05	Gaussian
14	equal	300	0.10	Gaussian
15	equal	300	0.15	Gaussian
16	equal	300	0.20	Gaussian
17	equal	500	0.01	Gaussian
18	equal	500	0.02	Gaussian
19	equal	500	0.03	Gaussian
20	equal	500	0.04	Gaussian
21	equal	500	0.05	Gaussian
22	equal	500	0.10	Gaussian
23	equal	500	0.15	Gaussian
24	equal	500	0.20	Gaussian
25	unequal	150	0.01	Gaussian
26	unequal	150	0.02	Gaussian
27	unequal	150	0.03	Gaussian
28	unequal	150	0.04	Gaussian
29	unequal	150	0.05	Gaussian
30	unequal	150	0.10	Gaussian
31	unequal	150	0.15	Gaussian
32	unequal	150	0.20	Gaussian
33	unequal	300	0.01	Gaussian
34	unequal	300	0.02	Gaussian
35	unequal	300	0.03	Gaussian
36	unequal	300	0.04	Gaussian
37	unequal	300	0.05	Gaussian
38	unequal	300	0.10	Gaussian
39	unequal	300	0.15	Gaussian
40	unequal	300	0.20	Gaussian
41	unequal	500	0.01	Gaussian
42	unequal	500	0.02	Gaussian
43	unequal	500	0.03	Gaussian
44	unequal	500	0.04	Gaussian
45	unequal	500	0.05	Gaussian
46	unequal	500	0.10	Gaussian
47	unequal	500	0.15	Gaussian
48	unequal	500	0.20	Gaussian

Table 20: *Factors of the simulation study with binary study variables*

no.	cluster size type	m	ρ	scale	π
1	equal	150	0.01	binary	0.05
2	equal	150	0.02	binary	0.05
3	equal	150	0.03	binary	0.05
4	equal	150	0.04	binary	0.05
5	equal	150	0.05	binary	0.05
6	equal	150	0.10	binary	0.05
7	equal	150	0.15	binary	0.05
8	equal	150	0.20	binary	0.05
9	equal	300	0.01	binary	0.05
10	equal	300	0.02	binary	0.05
11	equal	300	0.03	binary	0.05
12	equal	300	0.04	binary	0.05
13	equal	300	0.05	binary	0.05
14	equal	300	0.10	binary	0.05
15	equal	300	0.15	binary	0.05
16	equal	300	0.20	binary	0.05
17	equal	500	0.01	binary	0.05
18	equal	500	0.02	binary	0.05
19	equal	500	0.03	binary	0.05
20	equal	500	0.04	binary	0.05
21	equal	500	0.05	binary	0.05
22	equal	500	0.10	binary	0.05
23	equal	500	0.15	binary	0.05
24	equal	500	0.20	binary	0.05
25	unequal	150	0.01	binary	0.05
26	unequal	150	0.02	binary	0.05
27	unequal	150	0.03	binary	0.05
28	unequal	150	0.04	binary	0.05
29	unequal	150	0.05	binary	0.05
30	unequal	150	0.10	binary	0.05
31	unequal	150	0.15	binary	0.05
32	unequal	150	0.20	binary	0.05
33	unequal	300	0.01	binary	0.05
34	unequal	300	0.02	binary	0.05
35	unequal	300	0.03	binary	0.05
36	unequal	300	0.04	binary	0.05
37	unequal	300	0.05	binary	0.05
38	unequal	300	0.10	binary	0.05
39	unequal	300	0.15	binary	0.05
40	unequal	300	0.20	binary	0.05
41	unequal	500	0.01	binary	0.05
42	unequal	500	0.02	binary	0.05
43	unequal	500	0.03	binary	0.05
44	unequal	500	0.04	binary	0.05
45	unequal	500	0.05	binary	0.05
46	unequal	500	0.10	binary	0.05
47	unequal	500	0.15	binary	0.05
48	unequal	500	0.20	binary	0.05
49	equal	150	0.01	binary	0.25
50	equal	150	0.02	binary	0.25
51	equal	150	0.03	binary	0.25
52	equal	150	0.04	binary	0.25
53	equal	150	0.05	binary	0.25

Continued on next page...

no.	*cluster size type*	*m*	ρ	*scale*	π
54	equal	150	0.10	binary	0.25
55	equal	150	0.15	binary	0.25
56	equal	150	0.20	binary	0.25
57	equal	300	0.01	binary	0.25
58	equal	300	0.02	binary	0.25
59	equal	300	0.03	binary	0.25
60	equal	300	0.04	binary	0.25
61	equal	300	0.05	binary	0.25
62	equal	300	0.10	binary	0.25
63	equal	300	0.15	binary	0.25
64	equal	300	0.20	binary	0.25
65	equal	500	0.01	binary	0.25
66	equal	500	0.02	binary	0.25
67	equal	500	0.03	binary	0.25
68	equal	500	0.04	binary	0.25
69	equal	500	0.05	binary	0.25
70	equal	500	0.10	binary	0.25
71	equal	500	0.15	binary	0.25
72	equal	500	0.20	binary	0.25
73	unequal	150	0.01	binary	0.25
74	unequal	150	0.02	binary	0.25
75	unequal	150	0.03	binary	0.25
76	unequal	150	0.04	binary	0.25
77	unequal	150	0.05	binary	0.25
78	unequal	150	0.10	binary	0.25
79	unequal	150	0.15	binary	0.25
80	unequal	150	0.20	binary	0.25
81	unequal	300	0.01	binary	0.25
82	unequal	300	0.02	binary	0.25
83	unequal	300	0.03	binary	0.25
84	unequal	300	0.04	binary	0.25
85	unequal	300	0.05	binary	0.25
86	unequal	300	0.10	binary	0.25
87	unequal	300	0.15	binary	0.25
88	unequal	300	0.20	binary	0.25
89	unequal	500	0.01	binary	0.25
90	unequal	500	0.02	binary	0.25
91	unequal	500	0.03	binary	0.25
92	unequal	500	0.04	binary	0.25
93	unequal	500	0.05	binary	0.25
94	unequal	500	0.10	binary	0.25
95	unequal	500	0.15	binary	0.25
96	unequal	500	0.20	binary	0.25
97	equal	150	0.01	binary	0.50
98	equal	150	0.02	binary	0.50
99	equal	150	0.03	binary	0.50
100	equal	150	0.04	binary	0.50
101	equal	150	0.05	binary	0.50
102	equal	150	0.10	binary	0.50
103	equal	150	0.15	binary	0.50
104	equal	150	0.20	binary	0.50
105	equal	300	0.01	binary	0.50
106	equal	300	0.02	binary	0.50
107	equal	300	0.03	binary	0.50
108	equal	300	0.04	binary	0.50
109	equal	300	0.05	binary	0.50

Continued on next page...

no.	*cluster size type*	*m*	ρ	*scale*	π
110	equal	300	0.10	binary	0.50
111	equal	300	0.15	binary	0.50
112	equal	300	0.20	binary	0.50
113	equal	500	0.01	binary	0.50
114	equal	500	0.02	binary	0.50
115	equal	500	0.03	binary	0.50
116	equal	500	0.04	binary	0.50
117	equal	500	0.05	binary	0.50
118	equal	500	0.10	binary	0.50
119	equal	500	0.15	binary	0.50
120	equal	500	0.20	binary	0.50
121	unequal	150	0.01	binary	0.50
122	unequal	150	0.02	binary	0.50
123	unequal	150	0.03	binary	0.50
124	unequal	150	0.04	binary	0.50
125	unequal	150	0.05	binary	0.50
126	unequal	150	0.10	binary	0.50
127	unequal	150	0.15	binary	0.50
128	unequal	150	0.20	binary	0.50
129	unequal	300	0.01	binary	0.50
130	unequal	300	0.02	binary	0.50
131	unequal	300	0.03	binary	0.50
132	unequal	300	0.04	binary	0.50
133	unequal	300	0.05	binary	0.50
134	unequal	300	0.10	binary	0.50
135	unequal	300	0.15	binary	0.50
136	unequal	300	0.20	binary	0.50
137	unequal	500	0.01	binary	0.50
138	unequal	500	0.02	binary	0.50
139	unequal	500	0.03	binary	0.50
140	unequal	500	0.04	binary	0.50
141	unequal	500	0.05	binary	0.50
142	unequal	500	0.10	binary	0.50
143	unequal	500	0.15	binary	0.50
144	unequal	500	0.20	binary	0.50

Table 21: *Summary of the simulation study with two-stage equal probability cluster sampling*

ρ	m	$Var\left(\hat{\bar{y}}^{(HT)}_{(\text{clu2})\lceil \text{eq} \rceil \langle \bullet \rangle \{\bullet\}}\right)$	$Var\left(\hat{\bar{y}}^{(HT)}_{(\text{srs})\lceil \text{eq} \rceil \langle \bullet \rangle \{\bullet\}}\right)$	$\widehat{deff}$
0.01	150	0.000394	0.000324	1.2169
0.01	300	0.000349	0.000331	1.0550
0.01	500	0.000338	0.000330	1.0271
0.02	150	0.000444	0.000324	1.3685
0.02	300	0.000369	0.000339	1.0871
0.02	500	0.000351	0.000335	1.0472
0.03	150	0.000500	0.000327	1.5284
0.03	300	0.000401	0.000330	1.2142
0.03	500	0.000347	0.000340	1.0185
0.04	150	0.000552	0.000329	1.6772
0.04	300	0.000406	0.000334	1.2155
0.04	500	0.000360	0.000337	1.0709
0.05	150	0.000592	0.000328	1.8046
0.05	300	0.000437	0.000336	1.3017
0.05	500	0.000364	0.000330	1.1007
0.10	150	0.000877	0.000328	2.6749
0.10	300	0.000531	0.000328	1.6208
0.10	500	0.000399	0.000335	1.1927
0.15	150	0.001116	0.000323	3.4510
0.15	300	0.000639	0.000331	1.9312
0.15	500	0.000432	0.000332	1.3001
0.20	150	0.001396	0.000327	4.2707
0.20	300	0.000728	0.000329	2.2142
0.20	500	0.000465	0.000329	1.4146

Table 22: *Summary of the simulation study with two-stage unequal probability cluster sampling*

ρ	m	$Var\left(\hat{\bar{y}}^{(HT)}_{(\text{clu2})[\text{un}] \langle \bullet \rangle \{\bullet\}}\right)$	$Var\left(\hat{\bar{y}}^{(HT)}_{(\text{srs})[\text{un}] \langle \bullet \rangle \{\bullet\}}\right)$	$\widehat{deff}$
0.01	150	0.000422	0.000335	1.2609
0.01	300	0.000395	0.000326	1.2092
0.01	500	0.000388	0.000332	1.1669
0.02	150	0.000472	0.000332	1.4201
0.02	300	0.000421	0.000336	1.2526
0.02	500	0.000391	0.000327	1.1952
0.03	150	0.000541	0.000327	1.6548
0.03	300	0.000433	0.000335	1.2910
0.03	500	0.000390	0.000332	1.1760
0.04	150	0.000591	0.000334	1.7698
0.04	300	0.000460	0.000335	1.3712
0.04	500	0.000408	0.000325	1.2556
0.05	150	0.000633	0.000328	1.9305
0.05	300	0.000473	0.000335	1.4118
0.05	500	0.000401	0.000327	1.2284
0.10	150	0.000887	0.000332	2.6755
0.10	300	0.000562	0.000332	1.6909
0.10	500	0.000442	0.000330	1.3391

Continued on next page...

ρ	m	$Var\left(\bar{\tilde{y}}^{(HT)}_{(\text{clu2})\,[\text{un}]\,\langle\bullet\rangle\,\{\bullet\}}\right)$	$Var\left(\bar{\tilde{y}}^{(HT)}_{(\text{srs})\,[\text{un}]\,\langle\bullet\rangle\,\{\bullet\}}\right)$	$\widehat{deff}$
0.15	150	0.001165	0.000326	3.5734
0.15	300	0.000679	0.000325	2.0866
0.15	500	0.000470	0.000335	1.4031
0.20	150	0.001460	0.000330	4.4289
0.20	300	0.000763	0.000327	2.3314
0.20	500	0.000503	0.000334	1.5080

Table 23: *Summary of the distribution of selected estimators of ρ with equal probability cluster sampling*

ρ	m	Typ	rel. Bias	relMSE
0.02	150	AOV	−0.000631	0.003189
0.02	150	F	0.003831	0.003185
0.02	150	F2	−0.000631	0.003189
0.02	150	FR	−0.023157	0.003163
0.02	150	REML	−0.020906	0.003860
0.02	150	ML	−0.045763	0.003911
0.02	150	Laplace	−0.282794	0.007997
0.02	300	AOV	−0.002583	0.004942
0.02	300	F	0.001696	0.004934
0.02	300	F2	−0.002583	0.004942
0.02	300	FR	−0.021843	0.004926
0.02	300	REML	−0.242646	0.011248
0.02	300	ML	−0.266609	0.011368
0.02	300	Laplace	−0.732444	0.016462
0.02	500	AOV	−0.009672	0.007584
0.02	500	F	−0.005102	0.007588
0.02	500	F2	−0.009672	0.007584
0.02	500	FR	−0.027620	0.007579
0.02	500	REML	−0.691004	0.019870
0.02	500	ML	−0.703489	0.019832
0.02	500	Laplace	−0.875246	0.018344
0.05	150	AOV	−0.001280	0.002318
0.05	150	F	0.001165	0.002313
0.05	150	F2	−0.001280	0.002318
0.05	150	FR	−0.013617	0.002302
0.05	150	REML	−0.001298	0.002319
0.05	150	ML	−0.013634	0.002303
0.05	150	Laplace	−0.110931	0.002675
0.05	300	AOV	−0.002251	0.002679
0.05	300	F	0.000171	0.002677
0.05	300	F2	−0.002251	0.002679
0.05	300	FR	−0.011425	0.002673
0.05	300	REML	−0.002696	0.002718
0.05	300	ML	−0.011930	0.002716
0.05	300	Laplace	−0.252169	0.009986
0.05	500	AOV	0.000266	0.003676
0.05	500	F	0.002481	0.003672
0.05	500	F2	0.000266	0.003676
0.05	500	FR	−0.007649	0.003671
0.05	500	REML	−0.046791	0.007108
0.05	500	ML	−0.056151	0.007214
0.05	500	Laplace	−0.710494	0.035430
0.10	150	AOV	−0.001823	0.002355
0.10	150	F	0.000916	0.002353
0.10	150	F2	−0.001823	0.002355
0.10	150	FR	−0.010508	0.002343
0.10	150	REML	−0.001817	0.002355
0.10	150	ML	−0.010502	0.002344
0.10	150	Laplace	−0.143681	0.003958
0.10	300	AOV	−0.002085	0.002033
0.10	300	F	0.000297	0.002031
0.10	300	F2	−0.002085	0.002033
0.10	300	FR	−0.007778	0.002030

Continued on next page...

ρ	m	Typ	rel. Bias	relMSE
0.10	300	REML	−0.002086	0.002032
0.10	300	ML	−0.007779	0.002030
0.10	300	Laplace	−0.183996	0.004996
0.10	500	AOV	−0.000331	0.002298
0.10	500	F	0.001853	0.002296
0.10	500	F2	−0.000331	0.002298
0.10	500	FR	−0.004829	0.002296
0.10	500	REML	−0.000329	0.002298
0.10	500	ML	−0.004827	0.002296
0.10	500	Laplace	−0.282655	0.013233
0.20	150	AOV	−0.000870	0.002179
0.20	150	F	0.000631	0.002166
0.20	150	F2	−0.000870	0.002179
0.20	150	FR	−0.007263	0.002175
0.20	150	REML	−0.000870	0.002179
0.20	150	ML	−0.007263	0.002175
0.20	150	Laplace	−0.208661	0.010303
0.20	300	AOV	0.000269	0.001478
0.20	300	F	0.001048	0.001476
0.20	300	F2	0.000269	0.001478
0.20	300	FR	−0.003464	0.001476
0.20	300	REML	0.000272	0.001478
0.20	300	ML	−0.003461	0.001476
0.20	300	Laplace	−0.241457	0.012706
0.20	500	AOV	0.001145	0.001412
0.20	500	F	0.001694	0.001410
0.20	500	F2	0.001145	0.001412
0.20	500	FR	−0.001522	0.001410
0.20	500	REML	0.001144	0.001412
0.20	500	ML	−0.001523	0.001410
0.20	500	Laplace	−0.288254	0.017582

Table 24: *Summary of the distribution of selected estimators of ρ with unequal probability cluster sampling*

ρ	m	Typ	rel. Bias	relMSE
0.02	150	AOV	−0.002704	0.003265
0.02	150	F	0.001150	0.004013
0.02	150	F2	−0.003300	0.003261
0.02	150	FR	−0.025823	0.003237
0.02	150	REML	−0.017505	0.003712
0.02	150	ML	−0.042111	0.003740
0.02	150	Laplace	−0.096233	0.003120
0.02	300	AOV	−0.003497	0.004944
0.02	300	F	0.001380	0.006774
0.02	300	F2	−0.003821	0.004941
0.02	300	FR	−0.023081	0.004926
0.02	300	REML	−0.187324	0.009743
0.02	300	ML	−0.213137	0.009903
0.02	300	Laplace	−0.140420	0.004347
0.02	500	AOV	−0.002108	0.007763

Continued on next page...

ρ	m	Typ	rel. Bias	relMSE
0.02	500	F	−0.005210	0.012241
0.02	500	F2	−0.002324	0.007760
0.02	500	FR	−0.020284	0.007748
0.02	500	REML	−0.560319	0.018466
0.02	500	ML	−0.577361	0.018467
0.02	500	Laplace	−0.192726	0.006219
0.05	150	AOV	−0.003717	0.002351
0.05	150	F	−0.003188	0.002668
0.05	150	F2	−0.004296	0.002349
0.05	150	FR	−0.016619	0.002336
0.05	150	REML	−0.003579	0.002350
0.05	150	ML	−0.016213	0.002336
0.05	150	Laplace	−0.111969	0.002629
0.05	300	AOV	0.002295	0.002763
0.05	300	F	0.004785	0.003547
0.05	300	F2	0.001978	0.002761
0.05	300	FR	−0.007208	0.002751
0.05	300	REML	0.001727	0.002766
0.05	300	ML	−0.007710	0.002757
0.05	300	Laplace	−0.148322	0.003349
0.05	500	AOV	−0.003586	0.003674
0.05	500	F	−0.000176	0.005427
0.05	500	F2	−0.003796	0.003672
0.05	500	FR	−0.011706	0.003670
0.05	500	REML	−0.026883	0.005281
0.05	500	ML	−0.036284	0.005380
0.05	500	Laplace	−0.208443	0.004936
0.10	150	AOV	−0.001372	0.002416
0.10	150	F	0.001729	0.002581
0.10	150	F2	−0.001921	0.002414
0.10	150	FR	−0.010607	0.002402
0.10	150	REML	−0.001306	0.002389
0.10	150	ML	−0.010191	0.002376
0.10	150	Laplace	−0.144480	0.004014
0.10	300	AOV	−0.000502	0.002071
0.10	300	F	0.002070	0.002465
0.10	300	F2	−0.000801	0.002070
0.10	300	FR	−0.006498	0.002065
0.10	300	REML	−0.000555	0.002073
0.10	300	ML	−0.006411	0.002069
0.10	300	Laplace	−0.180694	0.004879
0.10	500	AOV	0.000909	0.002346
0.10	500	F	0.003058	0.003194
0.10	500	F2	0.000709	0.002345
0.10	500	FR	−0.003790	0.002342
0.10	500	REML	0.000900	0.002324
0.10	500	ML	−0.003731	0.002320
0.10	500	Laplace	−0.230817	0.007029
0.20	150	AOV	−0.001578	0.002277
0.20	150	F	0.000134	0.002327
0.20	150	F2	−0.002065	0.002276
0.20	150	FR	−0.008455	0.002274
0.20	150	REML	−0.001564	0.002219
0.20	150	ML	−0.008049	0.002216
0.20	150	Laplace	−0.211303	0.010542
0.20	300	AOV	0.000458	0.001638

Continued on next page...

ρ	m	Typ	rel. Bias	relMSE
0.20	300	F	0.001472	0.001806
0.20	300	F2	0.000191	0.001638
0.20	300	FR	−0.003542	0.001636
0.20	300	REML	0.000474	0.001618
0.20	300	ML	−0.003346	0.001616
0.20	300	Laplace	−0.243288	0.012970
0.20	500	AOV	−0.000528	0.001468
0.20	500	F	0.000183	0.001838
0.20	500	F2	−0.000705	0.001468
0.20	500	FR	−0.003371	0.001468
0.20	500	REML	−0.000537	0.001467
0.20	500	ML	−0.003283	0.001467
0.20	500	Laplace	−0.288872	0.017659

Table 25: *Correlations of estimates for binary variables in Spain – round 1*

	AOV	KEQ	KPR	UBE	FLC	MAK	PEQ	PGP	PPR	REML	ML
AOV	1.000										
KEQ	0.962	1.000									
KPR	1.000	0.962	1.000								
UBE	1.000	0.962	1.000	1.000							
FLC	1.000	0.962	1.000	1.000	1.000						
MAK	0.908	0.874	0.908	0.908	0.908	1.000					
PEQ	0.926	0.802	0.926	0.927	0.926	0.879	1.000				
PGP	0.907	0.873	0.908	0.907	0.907	1.000	0.879	1.000			
PPR	0.945	0.860	0.945	0.945	0.945	0.965	0.970	0.965	1.000		
REML	0.861	0.720	0.861	0.862	0.861	0.838	0.947	0.839	0.919	1.000	
ML	0.861	0.720	0.861	0.862	0.862	0.838	0.947	0.839	0.919	1.000	1.000

Table 26: *Correlations of estimates for binary variables in Spain – round 2*

	AOV	KEQ	KPR	UBE	FLC	MAK	PEQ	PGP	PPR	REML	ML
AOV	1.000										
KEQ	0.920	1.000									
KPR	1.000	0.920	1.000								
UBE	1.000	0.919	1.000	1.000							
FLC	1.000	0.920	1.000	1.000	1.000						
MAK	0.600	0.603	0.599	0.601	0.600	1.000					
PEQ	0.653	0.533	0.653	0.654	0.654	0.796	1.000				
PGP	0.598	0.601	0.597	0.599	0.598	1.000	0.796	1.000			
PPR	0.659	0.592	0.658	0.659	0.659	0.923	0.964	0.923	1.000		
REML	0.625	0.496	0.625	0.626	0.626	0.732	0.913	0.733	0.889	1.000	
ML	0.625	0.496	0.625	0.626	0.626	0.732	0.914	0.733	0.890	1.000	1.000

Table 27: *Correlations of estimates for binary variables in Spain – round 3*

	AOV	KEQ	KPR	UBE	FLC	MAK	PEQ	PGP	PPR	REML	ML
AOV	1.000										
KEQ	0.978	1.000									
KPR	1.000	0.978	1.000								
UBE	1.000	0.978	1.000	1.000							
FLC	1.000	0.978	1.000	1.000	1.000						
MAK	0.299	0.328	0.299	0.299	0.299	1.000					
PEQ	0.342	0.318	0.342	0.342	0.341	0.839	1.000				
PGP	0.295	0.324	0.295	0.296	0.295	1.000	0.839	1.000			
PPR	0.338	0.343	0.338	0.339	0.338	0.951	0.964	0.951	1.000		
REML	0.731	0.611	0.731	0.731	0.731	0.127	0.209	0.125	0.170	1.000	
ML	0.731	0.611	0.731	0.731	0.731	0.126	0.209	0.124	0.169	1.000	1.000

Table 28: *Correlations of estimates for binary variables in France – round 1*

	AOV	KEQ	KPR	UBE	FLC	MAK	PEQ	PGP	PPR	REML	ML
AOV	1.000										
KEQ	0.967	1.000									
KPR	1.000	0.967	1.000								
UBE	1.000	0.967	1.000	1.000							
FLC	1.000	0.966	1.000	1.000	1.000						
MAK	0.954	0.997	0.954	0.954	0.953	1.000					
PEQ	0.978	0.902	0.978	0.978	0.978	0.889	1.000				
PGP	0.954	0.996	0.954	0.954	0.954	1.000	0.889	1.000			
PPR	0.998	0.967	0.998	0.998	0.998	0.959	0.979	0.959	1.000		
REML	0.971	0.909	0.971	0.972	0.971	0.893	0.969	0.893	0.969	1.000	
ML	0.958	0.892	0.958	0.958	0.958	0.878	0.963	0.878	0.957	0.989	1.000

Table 29: *Correlations of estimates for binary variables in France – round 2*

	AOV	KEQ	KPR	UBE	FLC	MAK	PEQ	PGP	PPR	REML	ML
AOV	1.000										
KEQ	0.881	1.000									
KPR	1.000	0.881	1.000								
UBE	1.000	0.880	1.000	1.000							
FLC	1.000	0.881	1.000	1.000	1.000						
MAK	0.860	0.691	0.860	0.861	0.860	1.000					
PEQ	0.939	0.692	0.939	0.940	0.939	0.877	1.000				
PGP	0.859	0.688	0.859	0.859	0.858	1.000	0.877	1.000			
PPR	0.939	0.709	0.939	0.939	0.939	0.942	0.985	0.942	1.000		
REML	0.928	0.716	0.928	0.929	0.928	0.878	0.973	0.878	0.967	1.000	
ML	0.928	0.716	0.928	0.929	0.928	0.877	0.973	0.877	0.967	1.000	1.000

Table 30: *Correlations of estimates for binary variables in France – round 3*

	AOV	KEQ	KPR	UBE	FLC	MAK	PEQ	PGP	PPR	REML	ML
AOV	1.000										
KEQ	0.901	1.000									
KPR	1.000	0.901	1.000								
UBE	1.000	0.900	1.000	1.000							
FLC	1.000	0.901	1.000	1.000	1.000						
MAK	0.834	0.769	0.835	0.834	0.834	1.000					
PEQ	0.693	0.383	0.693	0.696	0.693	0.608	1.000				
PGP	0.833	0.767	0.834	0.833	0.833	1.000	0.609	1.000			
PPR	0.876	0.677	0.876	0.877	0.875	0.908	0.877	0.908	1.000		
REML	0.674	0.370	0.674	0.677	0.675	0.567	0.919	0.568	0.825	1.000	
ML	0.674	0.370	0.674	0.677	0.675	0.567	0.919	0.568	0.826	1.000	1.000

Table 31: *Correlations of estimates for binary variables in Poland – round 1*

	AOV	KEQ	KPR	UBE	FLC	MAK	PEQ	PGP	PPR	REML	ML
AOV	1.000										
KEQ	0.964	1.000									
KPR	1.000	0.964	1.000								
UBE	1.000	0.963	1.000	1.000							
FLC	1.000	0.964	1.000	1.000	1.000						
MAK	0.960	0.949	0.960	0.960	0.960	1.000					
PEQ	0.971	0.884	0.971	0.972	0.971	0.940	1.000				
PGP	0.959	0.948	0.959	0.959	0.959	1.000	0.940	1.000			
PPR	0.981	0.919	0.981	0.981	0.981	0.974	0.992	0.974	1.000		
REML	0.915	0.837	0.915	0.915	0.915	0.889	0.937	0.889	0.933	1.000	
ML	0.915	0.836	0.915	0.915	0.915	0.888	0.938	0.888	0.933	1.000	1.000

Table 32: *Correlations of estimates for binary variables in Poland – round 2*

	AOV	KEQ	KPR	UBE	FLC	MAK	PEQ	PGP	PPR	REML	ML
AOV	1.000										
KEQ	0.983	1.000									
KPR	1.000	0.983	1.000								
UBE	1.000	0.983	1.000	1.000							
FLC	1.000	0.983	1.000	1.000	1.000						
MAK	0.832	0.824	0.831	0.832	0.832	1.000					
PEQ	0.814	0.766	0.814	0.815	0.815	0.920	1.000				
PGP	0.830	0.822	0.830	0.831	0.831	1.000	0.920	1.000			
PPR	0.843	0.821	0.842	0.843	0.843	0.990	0.963	0.990	1.000		
REML	0.704	0.701	0.704	0.705	0.705	0.861	0.755	0.861	0.835	1.000	
ML	0.704	0.701	0.704	0.705	0.705	0.861	0.755	0.861	0.835	1.000	1.000

Table 33: *Correlations of estimates for binary variables in Poland – round 3*

	AOV	KEQ	KPR	UBE	FLC	MAK	PEQ	PGP	PPR	REML	ML
AOV	1.000										
KEQ	0.971	1.000									
KPR	1.000	0.971	1.000								
UBE	1.000	0.971	1.000	1.000							
FLC	1.000	0.971	1.000	1.000	1.000						
MAK	0.750	0.779	0.749	0.750	0.750	1.000					
PEQ	0.758	0.775	0.758	0.758	0.758	0.971	1.000				
PGP	0.748	0.778	0.748	0.749	0.748	1.000	0.970	1.000			
PPR	0.756	0.782	0.756	0.756	0.756	0.996	0.986	0.996	1.000		
REML	0.386	0.377	0.386	0.387	0.386	0.461	0.549	0.460	0.489	1.000	
ML	0.386	0.378	0.386	0.387	0.386	0.462	0.549	0.462	0.490	1.000	1.000

Zusammenfassung der Dissertation

In multinationalen Stichprobenerhebungen wie dem European Social Survey (ESS) werden in verschiedenen Ländern Stichproben erhoben um substanzielle Analysen zu ermöglichen. Ein zentrales Ziel, das es bei der Stichprobenplanung zu berücksichtigen gilt, ist die Vergleichbarkeit von Kennwerten zwischen Ländern und Ländergruppen. Die Erreichung dieses Ziels wird jedoch erschwert durch Unterschiede in den Stichprobendesigns zwischen den Ländern. Diese Unterschiede gehen auf divergierende Grundlagen für die Stichprobenziehung zurück. In einigen Ländern muss auf *komplexe*, z.B. mehrstufige Ziehungsverfahren zurückgegriffen werden während in anderen einfache oder geschichtete Zufallsauswahlen realisiert werden können. Im Rahmen der vorliegenden Dissertation wird jedes Stichprobenverfahren als komplex bezeichnet, das nicht eine einfache Zufallsauswahl von Personen darstellt. Ein häufig vorkommendes komplexes Auswahlverfahren ist die mehrstufige Auswahl. Ein Nachteil mehrstufiger Auswahlverfahren wird durch die räumliche Klumpung der befragten Personen hervorgerufen: Auf vielen interessierenden Merkmalen sind sich Befragte innerhalb eines Klumpens ähnlicher als zu allen anderen Elementen der Stichprobe. Dieser Umstand stellt eine Verletzung der Unabhängigkeitsannahme dar, die vielen Schätzern unterliegt. In der Folge wird die Varianz vieler Punktschätzer unterschätzt wenn trotzdem die herkömmlichen Formeln für die Varianzschätzung herangezogen werden (etwa $\frac{Var(y)}{n-1}$ als Varianzschätzer für den Mittelwert). Der Grad der Unterschätzung wird als *Designeffekt* bezeichnet und ist definiert als

$$deff = \frac{Var_c\left(\hat{\theta}\right)}{Var_{srs}\left(\hat{\theta}\right)}, \tag{1}$$

wobei $Var_c\left(\hat{\theta}\right)$ die Varianz des Schätzers $\hat{\theta}$ für den Populationsparameter θ unter dem gegebenen komplexen Stichprobendesign darstellt und $Var_{srs}\left(\hat{\theta}\right)$ die Varianz des selben Schätzers unter einfacher Zufallsauswahl. Ist $\hat{\theta}$ das Stichprobenmittel und werden durch einstufiges Auswahlverfahren m Primäreinheiten vom Umfang B ausgewählt und alle Elemente der Primäreinheiten erhoben, so lässt sich die obige Gleichung umformen zu

$$deff_{\text{one-stage}} = 1 + (B-1)\,\rho, \tag{2}$$

wobei ρ der Intralassen-Korrelationskoeffizient ist. Die Formulierung in Gleichung (1) wird als *Design-basierter* Ansatz bezeichnet, diejenige in Gleichung (2) als *Modell-basierter* Ansatz. Bei einer zweistufigen Auswahl, bei der innerhalb jedes ausgewählten Klumpens eine Unterauswahl von b Sekundäreinheiten uneingeschränkt zufällig ausgewählt wird, ist der Modell-basierte Designeffekt gegeben durch

$$deff_{\text{two-stage}} = 1 + (b-1)\,\rho \tag{3}$$

und im Falle unterschiedlicher Klumpengrößen mit einem Erwartungswert der Umfänge der Primäreinheiten von $\bar{b} = \frac{n}{m}$ durch

$$deff_{\text{two-stage}^*} = 1 + \left(\bar{b} - 1\right)\rho. \tag{4}$$

Werden den Elementen der Population durch ein Auswahlverfahren unterschiedliche Auswahlwahrscheinlichkeiten zugeordnet, so muss die dadurch entstehende zusätzliche Varianz des Schätzers (z.B. des gewichteten Mittelwerts) im Modell-basierten Ansatz durch den *Designeffekt aufgrund von ungleichen Auswahlwahrscheinlichkeiten*, $deff_p$, berücksichtigt werden. Der Design-basierte Ansatz berücksichtigt Designgewichtung direkt durch die Verwendung eines adäquaten Schätzers im Nenner von Gleichung (1).

Üblicherweise muss der Intralassen-Korrelationskoeffizient auf Basis der Stichprobendaten durch einen adäquaten Schätzer, $\hat{\rho}$, geschätzt werden. Für Variablen unterschiedlichen Skalentyps wurden verschiedene Schätzer für ρ vorgeschlagen. Diese werden in der vorliegenden Dissertation in Kapitel 4 zunächst vorgestellt und deren Qualität in Kapitel 5 anhand einer Monte-Carlo Simulation untersucht. Auch der Zähler sowie der Nenner des Ausdrucks in Gleichung (1) müssen anhand der Stichprobendaten geschätzt werden. Die für die Design-basierte Schätzung des Designeffekts verwendeten Schätzverfahren werden im Rahmen dieser Dissertation ebenfalls mit Hilfe einer Monte Carlo Simulation bewertet. Darüber hinaus wird untersucht, inwiefern sich Designeffekte und Interviewereffekte trennen lassen, bzw. deren gegenseitiger Einfluss abgeschätzt werden kann. Die in den Simulationen gefundenen Erkenntnisse werden anhand der Stichprobendaten ausgewählter Länder des ESS empirisch überprüft.

Ausbildungs- und Studienverlauf

Matthias Ganninger

geboren am: 22. Juli 1980 in Bretten

Ausbildung

09/1991 - 06/2000	Melanchthon-Gymnasium Bretten
10/2000 - 02/2002	Universität Konstanz, Fachbereich für Politik- und Verwaltungswissenschaft, Grundstudium Verwaltungswissenschaft
03/2002 - 10/2002	Bundesministerium für Wirtschaft und Technologie, Praktikum im Organisationsreferat
11/2002 - 11/2005	Universität Konstanz, Fachbereich für Politik- und Verwaltungswissenschaft, Hauptstudium Verwaltungswissenschaft mit Schwerpunkt Evaluation und Policy-Analyse
12/2009	Abschluss der Promotion an der Universität Trier, Fachbereich IV, Wirtschafts- und Sozialwissenschaften, Mathematik, Informatik und Wirtschaftsinformatik, Erstgutachter: Prof. Dr. Ralf Münnich, Zweitgutachter: PD Dr. Siegfried Gabler

Berufliche Tätigkeit

Seit 11/2005	Wissenschaftlicher Mitarbeiter bei GESIS – Leibniz-Institut für Sozialwissenschaften, Center for Survey Design and Methodology (CSDM), Mannheim

Der Designeffekt gewinnt in multinationalen Stichprobenerhebungen wie dem European Social Survey (ESS) zunehmend an Bedeutung. Im Rahmen einer eine ex ante Harmonisierung des effektiven Stichprobenumfangs kommen einerseits modellbasierte Verfahren zur Prognose eines erwarteten Designeffekts zum Einsatz. Bei der ex post Schätzung des Designeffekts bieten sich zudem auch designbasierte Verfahren an. Die vorliegende Arbeit stellt den design- und modellbasierten Schätzansatz gegenüber und bewertet deren Güte und Eignung im praktischen Einsatz. Diese Bewertung erfolgt zum einen auf Basis einer umfassenden Monte-Carlo Simulationsstudie, zum anderen werden Daten aus dem ESS benutzt.

The design effect is receiving increased attention in multi-national sample survey projects like the European Social Survey (ESS). On the one hand, model-based methods are applied for the prediction of expected design effects in order to ex ante harmonize the effective sample size of different samples. On the other hand, also design-based estimators can be used for the ex post estimation of design effects from sample data. This thesis compares the design-based and the model-based approach to design effects and evaluates their quality and applicability in real-world situations. This evaluation is based on a large-scale Monte-Carlo simulation study and on data from selected countries of the ESS.

gesis Leibniz-Institut für Sozialwissenschaften

ISBN 978-3-86819-010-6
ISSN 1869-2869
19,- EUR